以最优且精确的方式缩减最近邻训练集

Josiah Rohrer* Simon Weber†

arXiv:2302.02132v1 [cs.CG] 4 Feb 2023

摘要

在最近邻分类中，使用具有给定分类的点构成的训练集 P 来对 $\mathbb{R}^d$ 中的每个点进行分类：每个点的分类都与其在 P中的最近邻相同。最近，Eppstein[SOSA'22]提出了一种算法，用于检测相关训练点，即满足 P和 $P \setminus \{p\}$ 会产生不同分类的点 $p \in P$。我们研究如何找到最小基数约简训练集 $P' \subseteq P$，使得 P和 P'产生相同的分类。我们证明，若 P处于一般位置，则相关点集合就是这样一个最小基数约简训练集。此外，我们证明，对于可能退化的 P，当 $d = 1$时，寻找最小基数约简训练集的问题属于 P；当 $d \geq 2$时，该问题是 NP 完全的。

1 引言

尽管最近邻分类是最古老、最容易描述的分类技术之一，最近邻分类 [12] 至今仍是监督学习中广泛使用的方法。一个由$\mathbb{R}^d$ 中带有已知分类标签的数据点组成的训练集 P，用于对 $\mathbb{R}^d \setminus P$中的新点进行分类。点 q的分类与其在 P中的最近邻相同（若出现并列，则按某个固定规则打破平局，或者该点会获得多个分类）。

最近邻分类有许多变体，例如 k-最近邻 [12], ，其中一个点会获得其 k 个最近邻中占多数的分类；此外还有近似最近邻 [27]。本文仅考虑上述基本版本。

最近邻分类及其高效实现的需求，推动了《计算几何》中的许多概念的发展。Voronoi 图将 $\mathbb{R}^d$ 划分为最近邻相同的区域，因此这些区域的最近邻分类也相同 [4]。此后，人们又将其扩展为高阶Voronoi图 [4], ，以类似的方式描述具有相同 k 最近邻的区域。大量研究致力于高效计算 Voronoi 图 [14, 20, 32] ，以及点定位技术，以确定给定查询点 [20, 28]所在的Voronoi 图区域。对于高维数据，任何基于显式存储或计算训练集Voronoi 图的技术都不可行，因为维度为 d时， n 个点的Voronoi 图复杂度可达 $\Theta(n^{\lceil d/2 \rceil})$ [30]。针对中高维数据，人们开发了各种近似最近邻搜索方法，例如基于四叉树的数据结构 [3, 2, 9, 16] 和局部敏感哈希 [1, 13, 19, 24, 27]。这些方法避免了对 d的指数级依赖，但与朴素地遍历完整训练集来计算查询点的最近邻相比，性能提升仍然有限。

与其改进最近邻算法和数据结构，不如缩小训练集的规模，从而节省大量时间和存储空间。以无损方式缩小训练集（即不改变任何查询点的分类）的一种常见方法是移除所有非相关点。相关点（有时也称为边界点）是指移除它后会改变分类的点 [11]。可以证明，一次性移除所有非相关点后，得到的训练集所产生的分类与原始训练集相同。已有一系列算法被开发出来，用于高效计算相关点集。当前最优算法由 Flores-Velazco [17] 提出，它能在任意固定维度下于 $O(nk^2)$时间内找出相关点集，其中 k 为相关点的数量。该算法是对 Eppstein [15]算法的略微调整版本。

相关点集不一定是能够产生相同分类的最小训练点子集。事实上，甚至不能保证相关点集包含这样一个最小子集。图1对此作了说明。在介绍其相关点查找算法的论文中，Eppstein [15] 推测，在高维情况下，寻找这样一个能够产生相同分类的最小子集，要比寻找相关点困难得多。本文表明，并不需要高维，这个问题在维度为 $d \geq 2$时，对于二分类已经是 NP 困难的。

请注意，无损训练集缩减并非唯一被研究的方法。我们将在后文讨论其他方法

*苏黎世联邦理工学院数学系, rohrerj@student.ethz.ch

†苏黎世联邦理工学院计算机科学系, simon.weber@inf.ethz.chSimon Weber 获得瑞士国家科学基金会项目编号 204320 的资助。

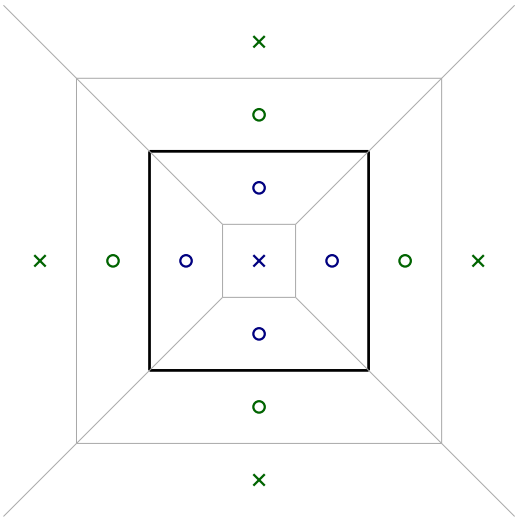


图1：一个训练集，其中相关点（圆圈）不包含唯一的最小基数约简训练集（叉号）。

这些方法允许（轻微地）改变诱导出的分类，详见第1.4节。

1.1 定义

定义 1 一个带标签点集（m, P, c）由一个整数 m、一个大小为 n的集合 $P \subset \mathbb{R}^d$ 以及一个分类函数 $c\colon\ P \to [m]$给出。我们称 c（p）为 p的标签。

我们用 $d(p, q)$表示两点 p, q之间的欧几里得距离。我们用 $\mathrm{nn}(q, P)$表示 q 在 P中的最近邻集合。若一个点集不含三点共线且不含四点共圆，则称其处于一般位置。

定义 2 一个带标签点集（m, P, c）诱导出最近邻分类 f：$\mathbb{R}^d \to 2^{[m]}$，其中

$$f(q) = \{c(p) \mid p \in nn(q, P)\}.$$

定义3 某个带标签点集的约简训练集 (m, P, c) 是满足以下条件的集合 $Q \subseteq P$：$(m, Q, c|_Q)$诱导出与(m, P, c)相同的最近邻分类。

定义4 判定问题 *d-MinNN* 是判断给定带标签点集(m, P, c) 是否存在一个至多包含 k 个点的约简训练集。

定义 5 $p \in P$ 是一个相关点，当$(m, P \setminus \{p\}, c|_{P\setminus\{p\}})$和$(m, P, c)$ 诱导出不同的最近邻分类时。我们用$\mathrm{rel}(P)$表示所有相关点组成的集合。

1.2 结果

现在我们可以陈述结果了。作为第一个结果，我们证明，对于处于一般位置的训练集，d-MinNN 可以轻松求解，因为相关点本身已经构成一个解。

定理1 对于 *d-MinNN* 的一个实例，若 P 处于一般位置，则 $\mathrm{rel}(P)$ 是唯一解。

如果不满足一般位置这一假设，则相关点集不一定是最优解。我们将证明，通常只有在维度为一时，才能找到基数最小的约简训练集；在其他情况下，该问题是NP完全的。

定理2 1-*MinNN* 属于P。

定理3 对于任意固定维度 $d \geq 2$，*d-MinNN* 是NP-完全的，即使对于二分类也是如此，也就是说，$m = 2$。

1.3 讨论

当数据点从带有（例如高斯）噪声的概率分布中独立采样时，所得数据集以概率1处于一般位置。因此，定理1表明，在实践中，当分类模型由来自噪声源的高精度数据训练而成时，使用 Eppstein [15] 或 Flores-Velazco [17] 的算法计算相关点，是以无损方式将训练集大小约简至最优值的一种高效方法。进一步约简训练集大小的唯一方法，是接受一些小的错误。尽管我们的结果并未表明近似训练集约简是困难的，但定理3表明，无法通过先对训练集进行“去噪”，再寻找最小基数约简训练集的方式高效地实现这一目标。

1.4 相关工作

关于最近邻训练集归约的大部分研究都聚焦于检测相关点。由于相关点的数量 k 预计远小于点的总数 n，寻找相关点的算法理想情况下应具有输出敏感性。Clarkson在 1994 [11]提出了首个此类算法。Bremner等人 [6] 改进了Clarkson针对具有两个标签的二维数据所提出的算法。最近，Eppstein [15] 提出了一种适用于任意维度和任意数量标签的算法，该算法基于简单的几何基元：计算欧几里得最小生成树和点集的极点。随后，Flores-Velazco [17] 证明可以省去欧几里得最小生成树这一步，从而得到适用于任意常数维度 d的 $O(nk^2)$算法。

最近邻训练集的有损归约，即通过归约使分类略有改变，在文献中通常称为最近邻凝缩 。凝缩中最常见的概念是一致子集，由*Hart*在中提出。一个一致子集是训练点的子集，它在原始训练集上产生相同的分类，但不一定对中的所有点都如此。已知，对于任意数量的标签，求最小基数一致子集是*NP*完全问题。选择性子集 [29] 是满足比一致子集更强条件的子集。具体而言，原始训练集中的每个点 p 到子集中具有相同分类的点的距离，都必须小于 p 到原始训练集中最近的不同分类点的距离。计算最小基数选择性子集同样是NP完全问题 [34]。Flores-Velazco和Mount [18] 提出了近似概念 α-一致性和 α-选择性，并证明寻找最小基数的 α-一致子集和 α-选择性子集不仅是NP 困难问题，而且要将其大小近似到超过特定近似因子也同样是NP 困难问题。由于这些NP 困难性结果，近期的大部分研究都聚焦于启发式方法，以保证所得子集的大小 [25]。

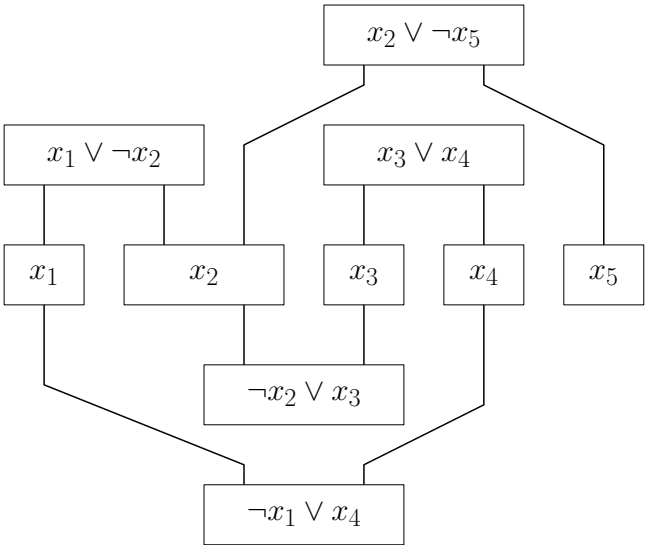


图2：定理3的证明中使用的变量-子句图嵌入类型。

到目前为止，我们讨论的只是通过取原始数据的子集来缩减训练集。当然，另一种选择是构造一个全新的训练集，使其产生的分类与原始数据（近似）相同，同时包含更少的点。Heath 和 Kasif [22] 指出，这种方法的精确版本行不通，因为他们证明，找出构造一个 Voronoi 图所需的最少点数（该图包含给定的多边形镶嵌作为子结构）是 NP 困难。这在一定程度上解释了为什么最近邻研究群体很少研究这种方法。

1.5 证明技术

定理1的证明非常直接。证明利用了这样一个观察：在每个缩减后的训练集 $Q \subseteq P$中，两个分类不同的区域之间的每一条 Voronoi墙 P 都必须位于某一对点的平分超平面上。如果 P处于一般位置，则任意两对点的平分超平面都不相同。

为证明定理2，我们给出从 1-MinNN 到区间图上求最大权独立集问题的归约，该问题可在多项式时间内求解 [23]。

我们对定理3中 NP 困难性的证明，与 Heath 和 Kasif 对寻找 Voronoi覆盖问题的 NP 困难性所作的证明类似，都是通过从平面3SAT归约得到的。这些证明有两个主要差异。一方面， d-MinNN 的解必须在所有 $\mathbb{R}^d$ 上具有相同的分类。相比之下，在 Voronoi 覆盖问题中，只需将一个固定的镶嵌作为子结构包含在 Voronoi 图中。这意味着，在训练集中引入额外的 Voronoi墙时，我们必须更加谨慎。另一方面， d-MinNN 的任何解都必须是训练集的子集，而在 Voronoi 覆盖问题中，则允许使用任意点。这使我们能够更好地控制可能解的结构，也更容易排除不需要的解。

我们的证明通过从问题$V\text{-}cycle\ max2SAT$进行归约来实现。该问题是 max2SAT 的一个变体，其中二分变量-子句图即使在添加一条经过与变量$x_1, \ldots, x_n$对应的顶点的哈密顿回路之后，仍保持平面性。Buchin 等人近期已证明该问题是 NP 困难的 [7]。该图的平面性保证我们能够高效地找到一种特定类型的图嵌入，如图2所示。随后，将每个对应于变量 x_i 的方框替换为一个变量构件，即一个带标签点集，其中有两个具有相同诱导分类的严格子集。在这两个子集中进行选择，表示变量 x_i的取值。然后，通过图的边上的通道传递该取值。最后，将每个对应于子句 C_j 的方框替换为一个子句构件，即一个带标签点集：当且仅当另外两个点中至少有一个已存在时，最小基数约简训练集的大小才会减少一个。这两个点属于传入通道，分别携带该子句中两个文字的取值。因此，当且仅当子句得到满足时，子句构件所需的点数才会减少一个。由此得到的带标签

点集的最小基数约简训练集大小因此使我们能够确定 V-cycle max2SAT 实例中可同时满足的子句的最大数量。

这一归约中的主要技术挑战是

(i) 避免构件之间发生不必要的相互作用，因为不同标签的点可能在空旷空间中相距很远，却仍会相互作用；以及

(ii) 确保归约生成的点集仅含多项式数量的点，且坐标大小为多项式级。

1.6 论文概述

我们在第 2 节证明定理1。随后，我们在第 3 节证明定理 2。最后，第 4 节专门用于证明定理3。

2 一般位置

本节旨在证明定理1：

定理1 对于一个 *d-MinNN* 实例，其中 P 处于一般位置，$\mathrm{rel}(P)$ 是唯一解。

由于我们已经知道，对于带标签点集（m， P， c），相关点集合 rel(P) 是一个约简训练集，因此只需证明任何约简训练集都必须包含所有相关点。

首先介绍几个定义。沃罗诺伊墙 是沃罗诺伊图中维度为 $d-1$的单元。这些墙将两个满维沃罗诺伊单元分隔开来。如果某面沃罗诺伊墙分隔的两个单元中，由 P 诱导的最近邻规则给出不同的分类，则称该墙属于决策边界。

观察1 对于由$Q \subseteq P$给出的任何约简训练集， P 中属于决策边界的沃罗诺伊墙必须是 Q中某个沃罗诺伊墙的子集，因为 Q 必须诱导出与 P相同的最近邻分类。

注意，每个沃罗诺伊墙 W 都是属于相邻沃罗诺伊单元的两个点的平分超平面的子集。由于 W 是 $d-1$维的，它唯一确定了该超平面。接下来我们将证明，在一般位置假设下，每个超平面至多可以是一个点对的平分超平面。

引理4 对于处于一般位置的任意点集 $Q \subset \mathbb{R}^d$， Q 中任意两个不同的点对都不具有相同的平分超平面。

证明。 反设 a， $b \in Q$ 和c， $d \in Q$ 具有相同的平分超平面。首先注意由于 $b \neq d$， a， b， c， d必定两两不同，因为 a， b和 a， d具有不同的平分超平面。

点 a、 b、 c、 d不可能共线，因为一般位置假设要求任意三点都不共线。由于 a、 b和 c、 d具有相同的平分超平面，直线 ab 和 cd 必然平行。因此， a、 b、 c、 d必然位于同一平面上。此外，它们必然是等腰梯形的顶点，即一个圆内接四边形 [31]。因此， a、 b、 c、 d共圆，这与一般位置假设相矛盾。□

根据 [15] 中的证明，我们知道每个相关点 $p \in$ rel(P) 都与某个标签为 $c(q) \neq c(p)$的点q 共享一面Voronoi墙，也就是说，这面墙是决策边界的一部分。根据观察1和引理4，我们知道 p， q 因此必须属于每个约简训练集。由于这对每个相关点 p都成立，因此由 Q 给出的任何约简训练集都必须包含集合rel(P)，定理1由此得证。

3 一维情形

本节中，我们给出一种多项式时间算法，用于在$\mathbb{R}^1$中找到一个最小基数约简训练集。换言之，我们证明定理2：1-MinNN $\in$P。

给定训练集(m， P， c)所诱导的分类将 $\mathbb{R}^1$ 分解为一组分类相同的开区间。我们将这些开区间之间的点集 $B = \{b_1, \ldots, b_{t-1}\}$ 称为决策边界点。首先，我们提出一些对任意约简训练集(m， Q， $c|_Q$)均成立的观察。第一，对于任意 $i \in [t]$， $Q \cap C_i \neq \emptyset$。第二，对于任意 $i \in [t-1]$， b_i 是 Q中与其距离最近的较大邻点和较小邻点之间的中点。最后，对于任意最小基数约简训练集，必须有 $Q \cap C_i \leq 2$。

直观而言，为了得到较小的约简训练集，我们必须找到训练集的一个子集，该子集通常在每个区间中只包含一个点，并且这个点同时参与定义两者。我们将此形式化为以下链的概念，如图 3 所示。

定义 6 k-链是集合 $Q := \{q_1, \ldots, q_k\} \subseteq P$，其中存在一个整数 i， 使得：
(i) 对任意 $j \in \{1, \ldots, k\}$, $q_j \in C_{i+j}$，以及 *(ii)* 对任意 $j \in \{1, \ldots, k-1\}$, $\frac{q_j + q_{j+1}}{2} = b_{i+j}$。~~于是我们称~~ P' 覆盖 边界点 $b_i, \ldots, b_{i+k-1}$。

如果区间 [min(Q),max(Q)] 和 [min(Q'), max(Q')] 不相交，则称两条链 Q, Q' 相容。因此，相容链覆盖互不相交的边界点集合。

现在可以看出，任何最小基数约简训练集都必须是两两兼容的链的并集。任何两两兼容的链集合都可以

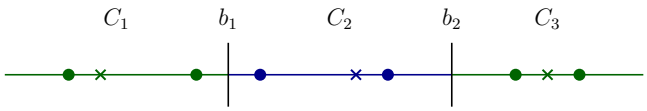


图 3： $\mathbb{R}^1$ 中的带标签点集，以及覆盖 b_1 和 b_2的 3-链（叉号）。

此外，还可以通过添加由相关点组成的 2-链，将其补全为约简训练集，以覆盖其余尚未覆盖的边界点。

如果我们的约简训练集是k_1-，……，$k_\ell-$条链的并集，则点的总数为 $2(t-1) - \sum_{i=1}^{\ell}(k_i-2)$，因为与仅由 2-链组成、包含 $2(t-1)$个点的朴素解相比，每条 k_i-链都能让我们在其覆盖的第一个和最后一个边界点之间的 k_i-2 个区间中节省一个点。

现在我们可以陈述完整算法。首先，我们计算边界点集和所有链的集合。注意，我们可以在 $O(n^2)$中轻松计算所有链的集合，并且它们的数量至多为 n^2。然后，我们将每条链 Q 与区间 $[\min(Q), \max(Q)]$关联起来。现在注意，两两兼容的链集是由这些区间给出的区间图中的独立集。我们为每条 2-链赋予一个极小权重 $\epsilon > 0$，并为每条 k-链赋予权重 $k-2$，其中 $k > 2$ 。最后，我们采用 [23] 的动态规划方法，在该图中寻找最大权独立集（*MWIS*）。该算法在线性于顶点数的时间内运行，因此在我们的情况下耗时为 $O(n^2)$。

所得独立集对应于一个包含极大的独立集，且其权重在所有此类集合中最大。因此，与之对应的链覆盖了所有边界点，而这些链的并集构成一个最小基数约简训练集。

4 NP 完全性

本节中，我们证明定理3。

定理3 对于任意固定维度 $d \geq 2$， *d-MinNN* 是NP完全问题，即使在二分类情形下也是如此，也就是说， $m = 2$。

我们给出 $d = 2$ 的 NP 困难性证明，因为任何 2-MinNN 实例都可以嵌入到 $\mathbb{R}^d$ 的一个二维子空间中，从而得到一个 d-MinNN 实例。该证明通过归约以下问题得到，而 Buchin 等人已证明该问题是 NP 困难的 [7][1]。

定义7 若合取范式公式 $\phi = C_1 \wedge \ldots \wedge C_b$ 包含变量 $x_0, \ldots, x_{a-1}$和一个整数 k，且每个子句 C_i 至多由两个文字组成，则它构成 V-cyclemax2SAT 问题的一个实例

并且图 $G_\phi = (V, E)$ 是平面的，其中

$$V = \{C_i \mid i \in [b]\} \cup \{x_i \mid 0 \leq i \leq a-1\},$$
$$E = \{(C_i, x_j) \mid x_j \in C_i\} \cup \{(x_i, x_{i+1 \bmod a}) \mid i \in [a]\}.$$

任务是判定是否存在变量 x_0、……、x_{a-1} 的一种赋值，使得 ϕ 中至少有 k 个子句得到满足。

注意，将 G_ϕ 中每个度为 2 的子句顶点 C_i 替换为一条边后得到的图 G'_ϕ 是哈密顿图，其哈密顿回路为 (x_0、……、x_{a-1}、x_0)。

图的 n-书嵌入 [5] 是指将图嵌入由 n 个半平面组成的空间中，这些半平面具有相同的边界直线 ℓ，且所有顶点都是 ℓ 上互不相同的点，每条边恰好与一个半平面的内部相交，并且除公共端点外，任意两条边都不相交。图的书厚度 $bt(G)$ 是满足 G 具有 n-书嵌入的最小 n 。

引理5（[5]） 对于每个平面（子）哈密顿图 G， bt（G）≤ 2。

这个引理的证明是构造性的；如果已知 G 中存在一条哈密顿路径，就能在多项式时间内完成构造。现在注意到，通过重新引入子句顶点并移除哈密顿路径，可以将 G'_ϕ 的 2-book 嵌入转换为 ϕ 的平面嵌入，其形式如图2所示。

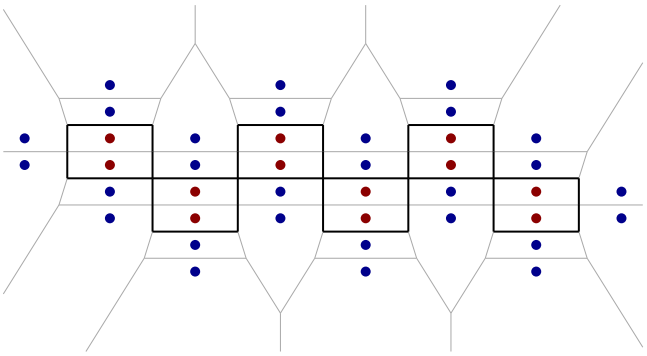
给定 ϕ 的这个嵌入，我们现在将构造一个带标签点集（2， P_ϕ， c_ϕ），用于编码 ϕ。为便于阅读，我们将这两个标签称为“红色”和“蓝色”。目标是找到某个点数 n'，使得存在一个至多包含 $n'-k$ 个点的约简训练集，当且仅当存在一种对 ϕ 中变量的赋值，能够满足至少 k 个子句。

为了将 ϕ 的嵌入转换为训练集，我们将说明如何使用变量构件编码变量，如何使用通道沿折线传递这些变量，以及如何使用子句构件编码子句。为确保各构件仅按预期相互作用，每个构件的设计都能将某个有界区域之外的所有点分类为蓝色。

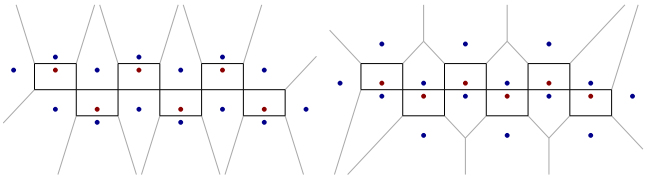
变量构件由图 4 所示的带标签点集构成。在该点集的每一半（上半部分和下半部分）中，只有两个严格子集（图 4a）会产生相同的分类：表示真值的外侧点（图 4b），以及表示假值的内侧点（图 4c）。请注意，这两个子集包含的点数相同。

图4a中最上方的6个点和最下方的6个点仅作为屏蔽点 使用，以确保构件外的每个点都被分类为蓝色。现在，变量构件的顶部和底部都连接了通道，以便将变量构件与上方或下方的子句构件相连。如果某个通道连接到一个变量为正出现的子句构件，该通道

[1]证明可在 arXiv 预印本 [8]的附录中找到。



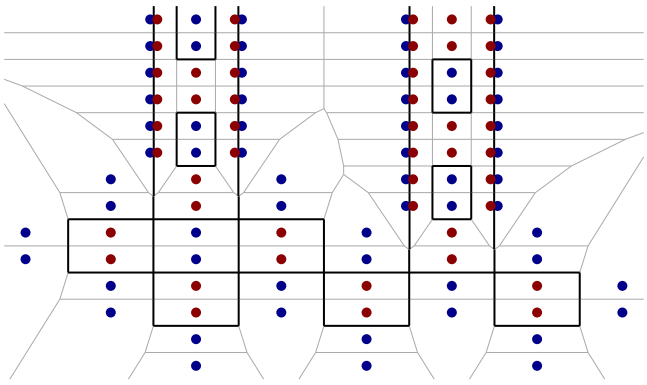
(a) 完整训练集。



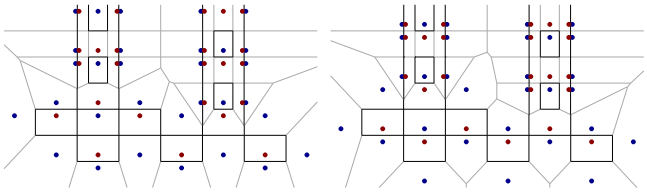
(b) “真”子集。 (c) “假”子集。

图 4：变量构件。

就连接到没有屏蔽点的列上。否则，变量为负出现，该通道就连接到有屏蔽点的列上。通道从不连接到最外侧的列；变量构件同一侧的通道之间始终至少留有两列空列。注意，通过添加更多重复的点模式，可以水平扩展变量构件，从而容纳任意数量的通道。图5展示了一个连接了两个通道的变量构件。



(a) 完整的训练集。

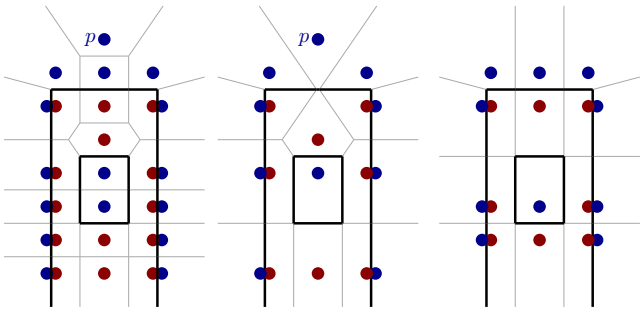


(b) “真”子集。 (c) “假”子集。

图5：顶部连接了两个通道的变量构件。左侧通道通向一个变量以肯定形式出现的子句，右侧通道通向一个变量以否定形式出现的子句。

通道所携带的真值将由子句构件通过检查在通道末端引入的点排列中是否存在某个点 p 来解读。该排列如图 6 所示。

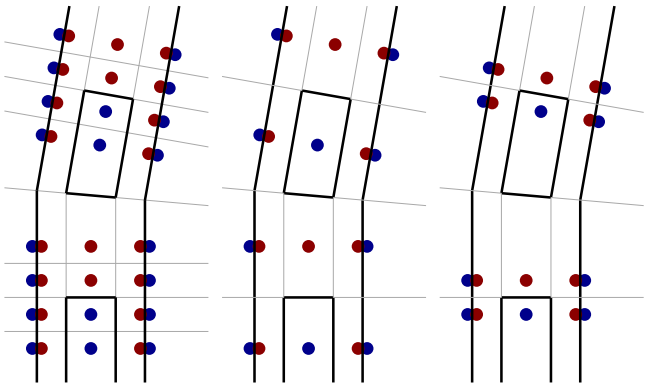
我们想简短地说明一下训练集中的点之间的距离。我们规定，变量构件（及通道）中相邻点之间的垂直距离为 1。变量构件中点之间的水平距离选为 3.2。这是经过深思熟虑的选择：一方面，水平距离过小会使图 6b中的点 p变得多余；另一方面，水平距离过大则会使一些用于生成通道左右垂直边界的红点和蓝点对变得多余。事实证明，3.2 这一取值能同时避免这两个问题。



(a) 所有点。 (b) “真”子集。 (c) “假”子集。

图6：通道的末端。当通道携带的值为真时，点 p 存在。

为了使我们能够连接所有变量构件和子句构件，通道需要具备更多灵活性。图 7 展示了如何在通道中实现某个固定小角度的弯曲。该角度是一个无理常数（接近 $10°$），其选取方式使得它的正弦和余弦均为有理数——



(a) 所有点。 (b) “真”子集。 (c) “假”子集。

图7：通道中 $\approx 10^\circ$ 的弯折。

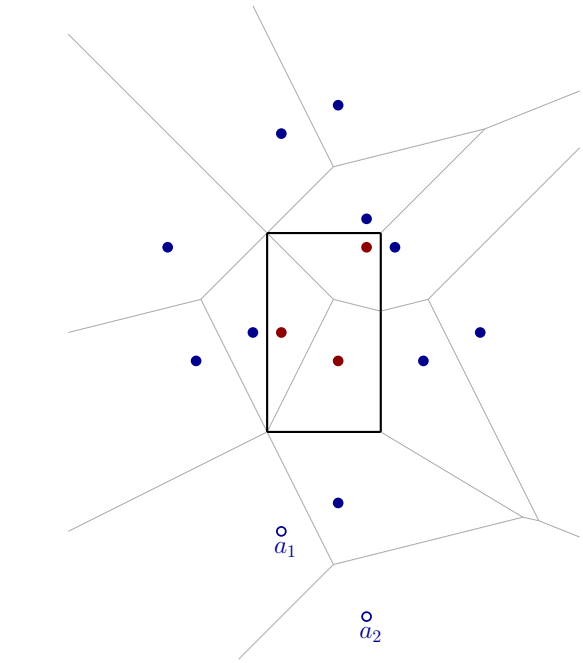
这确保了所有点的坐标都是有理数。我们还可以沿纵向（即重复模式的方向）拉伸通道，只需增大蓝色中心点所在行与红色中心点所在行之间的距离即可。借助弯曲和纵向拉伸的能力，我们可以使通道的端点位于所需位置，从而连接到子句构件。

图8展示了一个子句构件。它包含两个特殊点，分别标记为 a_1 和 a_2。这两个点是两条通道端点的位置，这些通道将相关变量的值传递给子句构件。请注意，为了使这些通道能够放置在图中且不影响子句构件的功能，a_1 和 a_2需要具有较小的垂直距离（例如 0.5）和较大的水平距离（例如 5）。子句构件也可以按照这些距离绘制，但为了清晰起见，图8中将水平距离和垂直距离调整为相等。附录 A 中附有一幅展示采用正确距离绘制的子句构件的图。

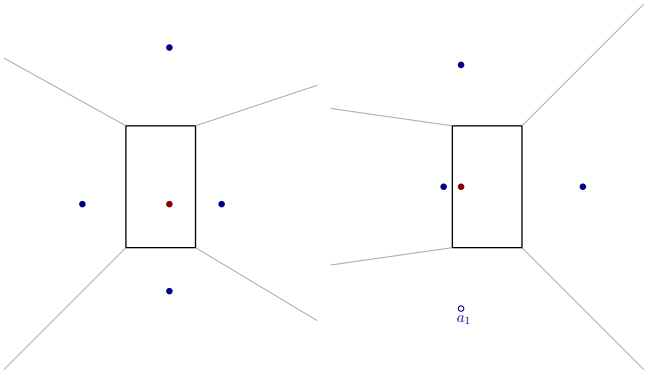
如果第一个（第二个）通道携带的值为真，则点 a_1 （a_2）属于约简训练集；否则不属于。子句构件的构造方式使得：当 a_1 和 a_2 均不存在时，需要 5 个点（图8b）；否则需要 4 个点（图8c 至图8e）。因此，当且仅当通道中现有的点所对应的变量赋值满足相应子句时，我们才能在子句构件中节省一个点。

现在，我们已经介绍了证明定理3所需的所有构件和技术

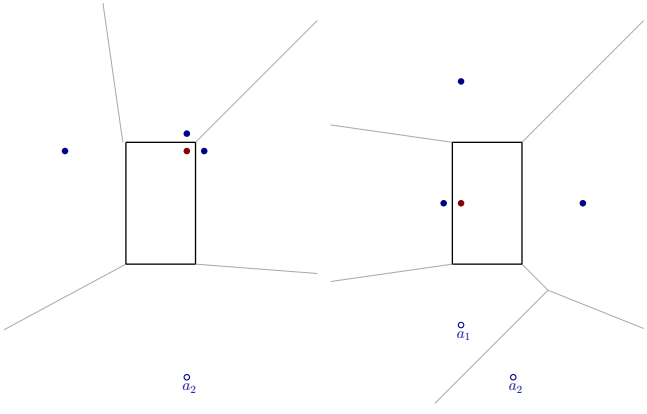
证明。 对于任意固定的 d，我们可以通过计算训练集和子集的沃罗诺伊图并比较分类结果，来验证给定的点子集是否为约简训练集。由于可以在多项式时间内计算 n 中的 Voronoi 图，这就证明了 NP 包含性。



(a) 完整的训练集。



(b) 两个变量均为假。 (c) 第一个变量为真。



(d) 第二个变量为真。 (e) 两个变量都为真。

图8：子句构件。

为了证明 NP 困难性，我们按照定义7从 V-cycle max2SAT 归约。给定一个 V-cycle max2SAT 实例 (ϕ, k)，我们可以在多项式时间内找到 G'_ϕ 的 2-book 嵌入。随后，将其转换为图2所示的 ϕ 的二分变量-子句图嵌入。每个对应变量的方框都替换为一个变量构件，每个对应子句的方框都替换为一个子句构件，而边则替换为通道（可能包含最多 2 个弯折和若干次拉伸）。由于该点集包含多项式数量的点，且所有点均可置于有理坐标上，所需描述位数至多为多项式数量，因此可以在多项式时间内构造该点集。

设 n_1 为与对变量真值的一种固定赋值相对应的所有变量构件和通道的约简训练集中的点数。设 n_2 为 $5{\cdot}b$（回顾一下，b 是 ϕ 中子句的数量）。现在我们将证明，存在一个大小至多为 $n_1 + n_2 - k$ 的最小基数约简训练集，当且仅当存在一种变量赋值，使得 ϕ 中至少有 k 个子句得到满足。

“若”方向显然成立：如果存在这样的变量赋值，就显然可以将其转化为大小合适的约简训练集，因为每个得到满足的子句构件只需要四个点。

对于“仅当”方向，我们证明，任何约简训练集都可以转化为与某个变量赋值对应的约简训练集，且点的数量不会增加。我们先考察通道。如果任何通道（包括其端点）使用的约简子集不是“真”或“假”子集，就必然使用了额外的点。我们将该子集改为与子句所连接的变量构件一侧所承载的真值相匹配的子集。这可能会使子句构件多用一个点，但也会使通道至少少用一个点。接下来，我们处理变量。如果任何变量构件使用的约简子集不是“真”或“假”子集，那么它至少有一侧（上半部或下半部）使用了全部点，而另一侧必定使用（a）全部点，或（b）仅使用“假”子集。令 m 为该构件中的列数。在情况（a）下，我们可以将该构件改为在两侧都使用“假”子集，并相应地切换所连接的通道和子句。每个相连的子句构件可能因此多用一个点，但变量构件至少能节省 $2m$ 个点，严格多于增加的点数，因为每一侧并非所有列都能被通道占据。在情况（b）下，我们将该构件改为在两侧都使用“真”子集。每个与先前为假的一侧相连的子句构件可能因此多用一个点，但变量构件至少能节省 m 个点，同样严格多于增加的点数。因此，我们得出结论：我们的训练集中至少存在一个最小基数约简训练集与某个变量赋值相对应，从而证明了我们的归约正确。□

致谢。 我们感谢 David Eppstein 在 SOSA’22 上所作的精彩报告，它启发了这项工作；也感谢 Bernd G¨artner 提供的宝贵建议。

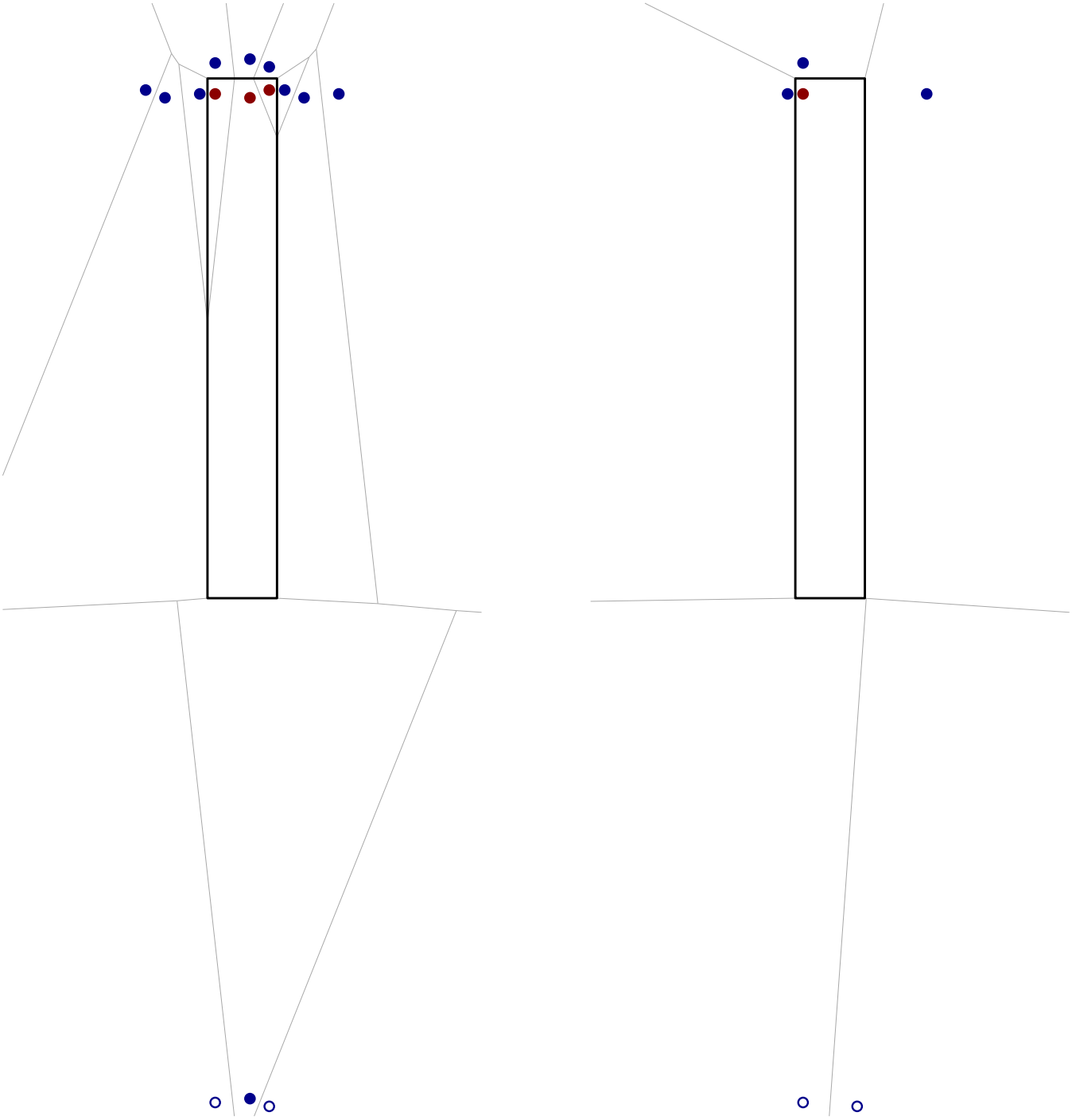
参考文献

[1] A. Andoni 和 P. Indyk。用于高维近似最近邻的近乎最优哈希算法。 *Commun. ACM*，51(1):117–122，2008年1月。[2] S. Arya、T. Malamatos 和 D. M. Mount。近似最近邻搜索的时空权衡。*J. ACM*，57(1)，2009年11月。[3] S. Arya、D. M. Mount、N. S. Netanyahu、R. Silverman 和 A. Y. Wu。固定维度下近似最近邻搜索的最优算法。 *J. ACM*, 45 6 891 923, 1998 ( ): – Nov 。[4] F. Aurenhammer、R. Klein 和 D.-T. Lee。 *Voronoi diagrams and Delaunay triangulations* 。世界科学出版公司，2013年。[5] F. Bernhart 和 P. C. Kainen。图的书厚度。 *Journal of Combinatorial Theory, Series B*，27 3 320 331, 1979 ( ): – 。[6] D. Bremner、E. Demaine、J. Erickson、J. Iacono、S. Langerman、P. Morin 和 G. Toussaint。用于计算最近邻决策边界的输出敏感算法。 *Discrete & Computational Geometry* 33 4 593 604, 2005 , ( ): – 。[7] K. Buchin 、 V. Polishchuk 、 L. Sedov 、 V. Roman 等。几何隔离路径与平面可满足性。载于 *36th International Symposium on Computational Geometry (SoCG 2020), June 22-26, 2020*年，第164卷，页码24–1，2020年。[8] K. Buchin、V. Polishchuk、L. Sedov 和 R. Voronov。几何隔离路径与平面可满足性。*arXiv preprint arXiv:1902.06471*，2019年。[9] T. M. Chan。重新审视近似最近邻查询。*Discrete & Computational Geometry*，20 3 359 1998 ( ): –373，Oct 。[10] B. Chazelle。任意固定维度下的最优凸包算法。 *Discrete & Computational Geometry*，10 4 377 409, 1993 ( ): – 。[11] K. L. Clarkson。更多输出敏感几何算法。载于*Proceedings 35th Annual Symposium on Foundations of Computer Science* 695 702 ，页码 – 。IEEE，1994年。[12] T. Cover 和 P. Hart。最近邻模式分类。 *IEEE Transactions on Information Theory*，13 1 21( ): –27，1967年。[13] M. Datar、N. Immorlica、P. Indyk 和 V. S. Mirrokni。基于 p 稳定分布的局部敏感哈希方案。载于 *Proceedings of the Twentieth Annual Symposium on Computational Geometry*，SCG ’04，页码253–262，纽约，美国，2004年。计算机协会。

[14] R. A. Dwyer。线性期望时间内构建高维沃罗诺伊图。*Discrete & Computational Geometry* 6 3 343 367, 1991 , ( ): – Sep 。[15] D. Eppstein。为最近邻分类寻找相关点。载于 *Symposium on Simplicity in Algorithms (SOSA)* 68 78 2022 , 页码 – 。SIAM, 。[16] D. Eppstein、M. T. Goodrich 和 J. Z. Sun。跳跃四叉树：用于多维点集的动态数据结构。*International Journal of Computational Geometry & Applications* 18 01 02 131 160, , (2008 n): – 。[17] A. Flores-Velazco。改进的最近邻分类相关点搜索。载于 S. Chechik、G. Navarro、E. Rotenberg 和 G. Herman 编，*30th Annual European Symposium on Algorithms (ESA2022)*，第244卷，*Leibniz International Proceedings in Informatics (LIPIcs)*，页码54:1–54:10，德国达格施图尔，2022年。达格施图尔城堡——莱布尼茨信息学中心。[18] A. Flores-Velazco 和 D. M. Mount。最近邻规则的核心集。载于 F. Grandoni、G. Herman 和 P. Sanders 编，*28th Annual European Symposiumon Algorithms, ESA 2020*，2020年，*Pisa, Italy (Virtual Conference)*，第173卷，*LIPIcs*，页码47:1–47:19。达格施图尔城堡——莱布尼茨信息学中心，2020年。[19] A. Gionis、P. Indyk 和 R. Motwani。通过哈希在高维空间中进行相似性搜索。载于 *25th VLDB Conference* 518 529, 1999 , 页码 – 。[20] L. J. Guibas、D. E. Knuth 和 M. Sharir。Delaunay 图和沃罗诺伊图的随机增量构造。*Algorithmica*，7(1):381–413，1992年6月。[21] P. Hart。压缩最近邻规则。*IEEE Transactions on Information Theory* 14 3 515 516, , ( ): 1968 – 。[22] D. Heath 和 S. Kasif。求解最小沃罗诺伊覆盖的复杂度及其在机器学习中的应用。*Computational Geometry*，3(5):289–305，1993年。[23] J. Y. Hsiao、C. Y. Tang 和 R. S. Chang。在区间图上求最大权二独立集的高效算法。*Information Processing Letters*，43 5 229 235, 1992 ( ): – 。[24] P. Indyk 和 R. Motwani。近似最近邻：迈向消除维数灾难。载于 *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*，STOC '98，页码604–613，纽约，美国，1998年。计算机协会。[25] N. Jankowski 和 M. Grochowski。实例选择算法比较 I：算法综述。载于 L. Rutkowski、J. H. Siekmann、R. Tadeusiewicz 和 L. A. Zadeh 编，*Artificial Intelligence and Soft Computing - ICAISC 2004*，页码598–603，柏林、海德堡，2004年。施普林格·柏林·海德堡。[26] K. Khodamoradi、R. Krishnamurti 和 B. Roy。两个标签的一致子集问题。载于 B. Panda 和 P. P. Goswami 编，*Algorithms and Discrete Applied Mathematics*，第131–142页，沙姆，2018年。施普林格国际出版公司。[27] T. Liu、A. Moore、K. Yang 和 A. Gray。对实用近似最近邻算法的研究。载于 L. Saul、Y. Weiss 和 L. Bottou 编，*Advances in Neural Information Processing Systems* 17 2004 , 第 卷。麻省理工学院出版社，。[28] F. P. Preparata 和 R. Tamassia。在凸空间胞腔复形中高效定位点。*SIAM Journal on Computing* 21 2 267 280, 1992 , ( ): –。[29] G. Ritter、H. Woodruff、S. Lowry 和 T. Isenhour。选择性最近邻判定规则算法。*IEEE Transactions on Information Theory*，21(6):665–669，1975年。[30] R. Seidel。d 维 Voronoi 图中面数的精确上界。载于*And Discrete Mathematics* 1990 *Applied Geometry* , 。[31] Z. Usiskin。*The classification of quadrilaterals: A study in definition* 2008 。信息时代出版社，。[32] D. F. Watson。计算 n 维德劳内剖分及其在 Voronoi 多胞体中的应用*。*The Computer Journal* 24 2 167 172, 1981 , ( ): –, 1 月。[33] G. Wilfong。最近邻问题。载于 *Proceedings of the Seventh Annual Symposium on Computational Geometry*，SCG '91，第 224–233 页，纽约，纽约州，美国，1991 年。美国计算机协会。[34] A. V. Zukhba。最近邻方法中原型选择问题的 NP 完全性。*Pattern Recognit. Image Anal.* 20 4 484 494, 2010 , ( ): –, 12 月。

子句构件

下方图9中的子句构件按比例绘制，其中通道 a_1 和 a_2 的两个连接点之间的水平距离为 5，垂直距离为 0.5。构件在垂直方向上拉伸是必要的，以确保当两个变量都为真时，分类仍然正确（图9b）。如果构件拉伸得不够，红点和 a_2 之间就会出现一面Voronoi墙。



(a) 完整的训练集。

(b) 两个变量都为真。

图9：子句构件及连接点之间的正确距离。