当真实值并非真实：对时间标注中人类偏差的建模

Taku Yamagata[1*]、Emma L. Tonkin[1*]、Benjamin Arana Sanchez[1],Ian Craddock[1],Miquel Perello Nieto[1],Raul Santos-Rodriguez[1],Weisong Yang[1],Peter Flach[1]布里斯托大学，英国
{taku.yamagata,e.l.tonkin,benjamin.aranasanchez,ian.craddock,miquel.perellonieto,enrsr,ws.yang,peter.flach}@bristol.ac.uk

arXiv:2302.02706v1 [cs.LG] 6 Feb 2023

摘要——**在监督学习中，低质量的标注会导致分类和检测模型表现不佳，同时也会使评估变得不可靠。这一点在时间数据中尤为明显，因为标注质量会受到多种因素的影响。例如，在事后自我报告日常活动时，认知偏差是最常见的影响因素之一。特别是，在活动结束后报告其开始时间和持续时长，可能会受到个人时间感知所引入的偏差影响，也会因时间取整而产生不精确和粒度不足的问题。本文提出一种对时间标注中的人类偏差进行建模的方法，并主张使用软标签。在合成数据上的实验结果表明，对于多种指标，软标签能够更好地近似真实值。我们还在一个真实的日常活动数据集上展示了该方法。**

I. 引言

用于检测和监测日常人类活动（例如运动、烹饪、卫生和睡眠）的系统，其开发有赖于准确标注的数据。事件日志记录方法（参与者自我报告事件）常用于自我标注，但存在若干局限。例如，其可靠性有限：Möller 等人 [1] 发现，在一项为期 6 周的智能手机使用研究中，受试者报告的事件至多仅占检测到事件的 70%，这可能是由于健忘。故意错报也可能影响某些活动的记录，例如健身 [2]。由于对任务进行同步标注通常需要中断参与者的活动，因此参与者很可能不会同步记录事件，以保持目标专注并避免过度切换任务 [3]；此外还有实际原因，例如，在吸尘或淋浴等任务期间，输入或写下标注并不现实。因此，就事件记录、事件开始/结束时间 和事件持续时间而言，参与者提供的标注可能不可靠且不完整。

时间估计中的不确定性反映了参与者使用粗略时间单位来描述时间和持续时间（例如，某个事件大约在某一时间开始，以一定的粒度进行估计[“大约在半点”]，并偏好“典型”时间 [4], [5], 同时以一定的持续时间进行估计）。语言学领域已对事件持续时间的自我报告进行了研究（通常通过大型语料库[6]），音乐领域也有相关研究 [7]。一天中的时间估计精确率反映了参与者的时间意识。时间感知十分复杂且具有弹性 [8]，并会根据年龄等人口统计特征而变化 [9] 以及情境（例如一天中的时间 [10]）而变化。还应注意，时间意识受损是各种诊断中的临床特征，例如阿尔茨海默病 [11]。目前，分析参与者标注的传感器数据时，尚未广泛对参与者时间标注的特征进行建模。

总之，尽管深入讨论超出了本文的范围，但除非有电子时间戳的支持，参与者对活动开始时间、结束时间和持续时间的记录通常并不精确；我们在第2节中给出了一个例子。这一质量问题可以与Kwon等人提出的“标签抖动” [12]相比较。然而，事后标注（例如对视频的标注）中的抖动幅度可能仅为几分之一秒。正如我们在第2节中所见，参与者对活动时间的估计具有较大的不确定性。关于评估标签质量 [13]以及从弱标签 [14] 中学习的大量研究，都源于事后标注；在这种标注方式下，可以比较多标注者表现（例如通过统计度量），并据此构建概率分布。

本文探讨了一种利用参与者贡献的标签来描述时间序列标注任务时，对时间不确定性进行建模的方法，并初步探索了该方法对特定应用（例如评估预测结果、训练模型）的适用性；全文组织如下。第II节介绍一个用于阐明问题的案例研究。第III节介绍我们估计标注模糊性并计算软标签的方法。第IV节使用人工数据集和真实世界数据集展示软标签的特征。第V节对我们的方法进行批判性讨论。最后，第VI节总结全文并提出可能的未来工作。

II. 案例研究：SRM-17 日常生活活动记录

在 SPHERE 100户项目中，研究人员鼓励部分参与者通过经过修改的社交节律量表（SRM-17）提供每日标注。该量表旨在衡量日常生活规律性 [15],，由参与者自行填写，最长持续七天，最终获得约 100 天的数据。对这一小型数据集的分析展示了引言中提到的许多特征。以淋浴/沐浴事件为例，如图 1 所示，参与者提供的标注集中在每小时内的特定时刻，例如整点或半点。数据往往精确到最接近的五分钟。这并非参与者所收到指示的要求；他们

*这些作者对本研究作出了同等贡献。

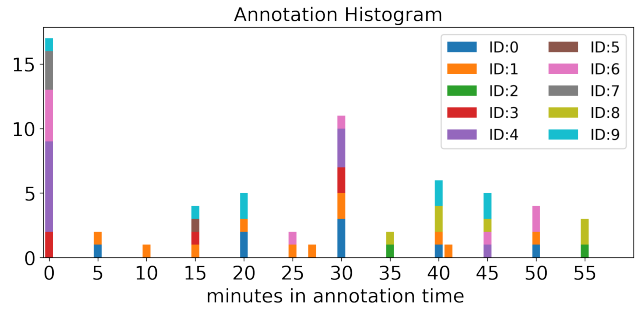

图 1：SPHERE100-Homes 数据集中 SRM-17 淋浴/沐浴标注中发现的粒度模式直方图，按标注者着色。

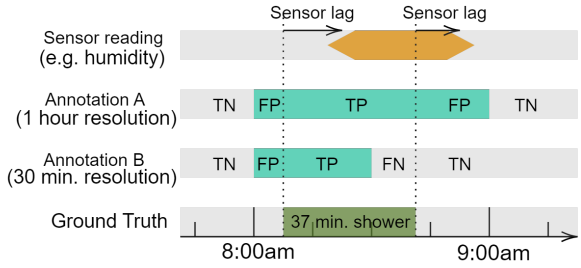

图 2：标注过程中引入的偏差示例。底部的图显示了此人洗澡 37 分钟的时间。上方的图显示了一种可能的标注**B**，估计此人在早上 8 点洗澡约 30 分钟；再上方的图则显示了另一种标注**A**，估计此人在早上 8 点洗澡一小时。每种标注都指出了与真实值比较后产生的真正例（和假正例）以及真负例（和假负例）。

并未得到有关时间报告精确率的具体指导；研究人员原本预计参与者会根据手表或时钟上的时间进行报告。尽管如此，在临床领域，指示中规定较低的时间报告精确率十分常见。例如，常用于评估帕金森病患者功能状态的 Hauser 家庭日记要求参与者每半小时填写一次日记 [16]。因此，我们可以设想，在处理临床数据时，数据分析人员可能会遇到以下两种情况之一：标注报告受到日记填写频率 的限制（例如“过去半小时内”）；或受到参与者选择的时间近似粒度 的限制（例如精确到最接近的五分钟、半小时等）。

虽然我们希望评估基于传感器数据做出的预测结果与参与者提供的标注之间的相关性，但传感器数据本身可能无法提供准确的时间信息。活动和事件预测依赖于信号，而这些信号会因住宅的一些偶然特征而变化，例如，烹饪产生的热量测量值会因炉灶类型、位置、响应时间和传感器位置而异 [17]。因此， [17] 会应用“热滞后”参数，以考虑检测到温度升高所需的时间。同样，用于预测淋浴或沐浴事件发生时间的湿度传感器也会出现“湿度滞后”，这是由浴室家用电器（例如淋浴设备、浴缸）的特性、布局（例如封闭式淋浴间、浴帘，或浴室内没有隔断）以及建筑结构（例如房间大小、窗户位置和通风方式，如排气扇）造成的。

III. 方法

我们提出分两个阶段处理时间序列标注中的不确定性。第一步，我们使用贝叶斯方法预测每位标注者采用的时间分辨率（模糊性）。第二步，根据预测的标注时间分辨率生成软标签。考虑如何使用软标签评估目标模型（即如何使用软标签）也很重要，但这超出了本文的讨论范围。

A. 预测标注者的习惯与标注的不确定性

首先，我们为标注时间分辨率引入预定义类别（c_n）。每个类别都有一组可能的标注值（$\mathbb{C}_n$），用于表示属于该类别的人倾向于如何进行标注

。例如，如果我们将第 1 类（c_1）定义为 30 分钟的时间分辨率，那么可能的标注时间就是 0 分钟和 30 分钟（$\mathbb{C}_1 = \{0, 30\}$）。我们通过本文中的以下五个类别来演示我们的方法。

g p p
c_1 : 30 mins. res. $\mathbb{C}_1 = \{0, 30\}$
c_2 : 15 mins. res. $\mathbb{C}_2 = \{0, 15, 30, 45\}$
c_3 : 10 mins. res. $\mathbb{C}_3 = \{0, 10, 20, 30, 40, 50\}$
c_4 : 5 mins. res. $\mathbb{C}_4 = \{0, 5, 10, 15, \ldots, 55\}$
c_5 : 1 mins. res. $\mathbb{C}_5 = \{0, 1, 2, 3, 4, \ldots, 59\}$

Th我们引入两个随机变量d 一个表示标注者的习惯（H_j），用于指示标注者倾向于采用哪种时间分辨率类别；另一个是实际用于每个标注的类别（$C_{j,\,i}$）。索引 j 和 i 分别表示标注者和标注的索引。$C_{j,\,i}$ 针对每个标注定义，而 H_j 针对每个标注者定义。现在，我们介绍用于推导这两个随机变量后验概率（H_j 和 $C_{j,\,i}$）的方法。为简单起见，在本文其余部分中省略标注者索引 j，因为以下过程可以针对每个标注者分别重复。下面给出符号说明。

D $= \{d_i\}_{i=1}^{N}$ 标注集合N 标注数量d_i^* 某个标注的真实值 d_iH 标注者的习惯
C_i 第 i 个标注类别c_n H 和 C_i 的类别
$|c_n|$ 类别 c_n 中的标注数量T_{c_n} 类别 c_n 对应的时间分辨率N_c 类别数量

我们假设第 i个标注（d_i）由C_i 和 d_i^*生成，而 C_i 由 H生成。图3展示了这些变量之间关系的概率图模型。为了推断标注者的 H 和标注的 C_i，我们根据标注 D 计算它们的后验概率，即 $P(H|\mathcal{D})$和 $P(C_i|\mathcal{D})$。

$$P(H|\mathcal{D}) = \frac{P(H)}{P(\mathcal{D})} \prod_{i=1}^{N} \sum_{C_i} P(C_i|H) P(d_i|C_i) \qquad (1)$$

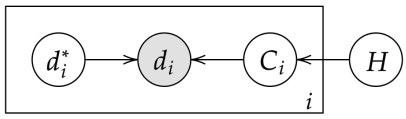

图3：描述标注相关变量之间关系的概率图模型。我们假设标注 d_i 由真实值 d_i^*和标注时间分辨率 C_i生成。标注时间分辨率取决于标注者的习惯 H。索引 i 是标注的索引。

公式1展示了如何计算 $P(H|\mathbb{D})$。详细推导见附录A。项 $P(d_i|C_i)$表示在给定标注类别C_i的情况下，得到标注 d_i 的概率。这里，我们假设 $P(d_i^*)$服从均匀分布；因此将其定义为公式2。

$$P(d_i|C_i) = \begin{cases} \frac{1}{|C_i|}, & \text{if } d_i \in \mathbb{C}_i \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

其中， $|C_i|$ 是类别 C_i中的标注数量。例如，当 $C_i = c_1$时，$|C_i| = 2$ 作为 c_1 有两个成员$\mathbb{C}_1 = \{0, 30\}$。因此，在公式2中，当$C_i = c1$ 时， $d_i = 0$ 或30的概率为0。5。术语 $P(C_i|H)$表示当标注者的习惯为 H时，使用标注类别 C_i 的概率。本文使用如下的 $P(C_i|H)$：

$$P(C_i|H) = \begin{cases} 1-\delta, & \text{if } C_i = H \\ \frac{\delta}{N_c - 1}, & \text{otherwise} \end{cases} \tag{3}$$

其中， $\delta \in [0, 1]$ 是选择不同于标注者习惯 H的标注类别的概率。本文取值为 $\delta = 0$。1。关于选择 δ值的进一步讨论见第V节。我们假设 H 的先验为均匀分布，因为我们不假设关于标注者习惯的任何先验知识。现在，我们计算 C_i的后验，可推导如下（详细推导同样见附录B）：

$$P(C_i|\mathcal{D}) = \sum_H \frac{P(d_i|C_i)P(C_i|H)}{P(d_i|H)} P(H|\mathcal{D}), \tag{4}$$

其中，我们可以根据公式2计算 $P(d_i|C_i)$，根据公式3计算 $P(C_i|H)$，根据公式1计算$P(H|D)$，并计算 $P(d_i|H) = \sum_{C_i} P(d_i|C_i)P(C_i|H)$。计算出后验后，我们用它们生成软标签。我们可以为所有类别准备软标签，并根据后验将它们组合起来（完全贝叶斯方法）。本文对C_i 采用最大后验（MAP）估计，并据此生成软标签。

B. 生成软标签

时间序列数据的标注是通过记录事件的开始和结束时间生成的。我们根据上一小节检测到的类别 C_i ，为这些时间点引入概率分布，并计算软标签，以表示事件在给定时间发生的概率。我们对真实的开始和结束时间 d_s^* 和d_e^* 采用均匀分布，因为我们假设标注是通过对实际时间取整生成的。（同样，研究不同分布是未来工作中一个有趣的方向。）这些分布在推断出的标注区间内均匀分布，且其中心位于标注的时间点。例如，如果开始时间的标注类别 C_s 为 c_3，则其概率分布的范围为$d_s - 5$ 和 $d_s + 5$ ，因为类别 c_3 的标注间隔为 10 分钟，其中 d_s 和 d_e 分别是开始和结束时间的标注。

$$\begin{aligned} p(d_s^*) &= \mathrm{U}(d_s - T_{C_s}/2, d_s + T_{C_s}/2), \\ p(d_e^*) &= \mathrm{U}(d_e - T_{C_e}/2, d_e + T_{C_e}/2), \end{aligned} \tag{5}$$

其中， T_{C_s}和 T_{C_e}是推断的标注类别 C_s 和 C_e的标注时间区间。然后，我们通过对公式5进行积分，计算事件在时间 t 已经开始 $P(d_s^* \le t)$但尚未结束$P(d_e^* > t)$的概率。

$$\begin{aligned} P(d_s^* \le t) &= \int_{-\infty}^{t} p(d_s^*)\, dd_s^*, \\ P(d_e^* > t) &= \int_{t}^{+\infty} p(d_e^*)\, dd_e^*. \end{aligned} \tag{6}$$

最后，我们通过将 $P(d_s^* < t)$与 $P(d_e^* > t)$相乘，计算软标签（$P^{(\text{label})}(t)$），即事件在时间 t 发生的概率。这里，我们假设开始和结束时间在统计上相互独立。这简化了后续推导，也有助于阐明我们的方法。这是一个很强的假设，在某些情况下并不成立。例如，如果开始和结束时间相对于标注时间分辨率而言相距较近，那么我们就需要考虑开始和结束时间之间的依赖关系——结束时间必须晚于开始时间。我们希望未来扩展模型以支持这种情形。

$$\begin{aligned} P^{(\text{label})}(t) &= P(d_s^* \le t \wedge d_e^* > t) \\ &= P(d_s^* \le t) P(d_e^* > t). \end{aligned} \tag{7}$$

图4展示了采用均匀分布计算上述软标签的过程。首先，根据给定的起止时间标注（d_s 和 d_e）以及估计的时间分辨率，假设实际起止时间（d_s^* 和 d_e^*）的概率分布。然后，计算每个时间槽中事件已经开始和已经结束的概率。最后，计算事件已经开始但尚未结束的概率，并将其作为软标签。

IV. 实验

本节通过人工示例和真实世界数据集比较硬标签与软标签。硬标签的值为真或假，并将起止时间标注之间的时间槽设为真。人工示例有助于我们在受控环境中展示软标签的特征。我们使用真实世界数据集展示实际用例。

*这些概率必须以 C_i 和 d_i 为条件，例如 $p(d_s^*|C_s、\ d_s)$。为简洁起见，我们在此省略条件。

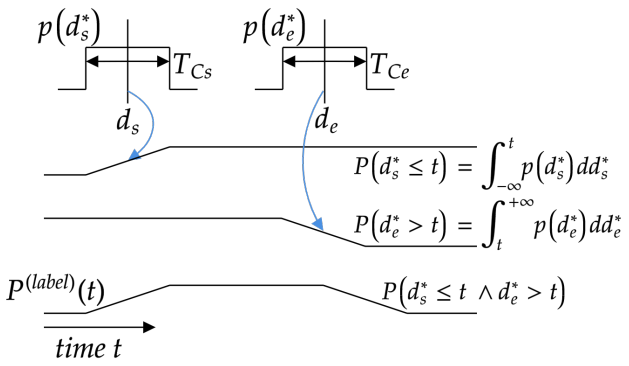

图4：时间序列数据的软标签。给定起止时间标注（d_s 和 d_e）及估计的时间分辨率，我们假设实际起止时间（d_s^* 和 d_e^*）的概率分布——见图上方。然后，我们计算每个时间槽中事件已经开始和已经结束的概率。最后，我们计算每个时间槽中事件已经开始但尚未结束的概率，并将其作为软标签。

A. 简单任务

首先，我们通过一个简单的合成示例比较软标签和传统硬标签。该示例由一系列包含开始和结束事件的任务组成（例如睡觉、烹饪或淋浴）。在这里，我们随机生成真实的开始和结束时间（真实值），然后根据给定的时间分辨率（例如 5、10、15 和 30 分钟）对时间进行取整，以生成标注。我们将取整后的开始和结束时间视为提供的硬标签，并按照第III节所述应用所提出的方法来获得软标签。最后，我们通过将硬标签和软标签与真实值进行比较来评估它们。图5展示了真实值相对于硬标签和软标签的均方误差（MSE），X轴表示标注的时间分辨率。MSE是在真实开始和结束时间周围±15 分钟的范围内测量的。结果表明，随着时间标注的分辨率间隔增大，硬标签的误差逐渐大于软标签。这清楚地表明，所提出的软标签受到的惩罚小于硬标签。

接下来，我们对 F1分数进行类似的分析，因为在大多数检测场景中，正类比负类更重要。图6左侧的图展示了以标注时间分辨率为X轴的结果。这表明硬标签优于软标签（F1分数更高）。这与均方误差的结果相反，也有悖直觉：软标签的取值介于0和1之间，能够准确反映模糊程度，而硬标签的取值只能是0或1。F1分数（或任何基于混淆矩阵的指标）会因取值介于0和1之间（如软标签）而受到惩罚。此外，这也是硬标签的最佳情况，因为标注是通过取整生成的。因此，预期的真实值与硬标签相符。我们希望探索不会惩罚预测或真实值模糊性的评估指标

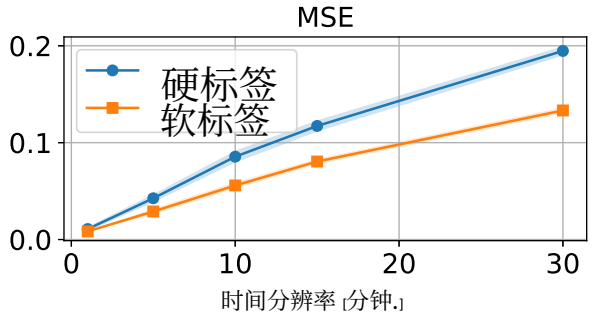

图5：硬标签和软标签相对于真实值的均方误差。X轴为标注时间分辨率。均方误差是在真实值开始和结束时间 ±15分钟范围内测量的。结果表明，软标签优于硬标签（均方误差更小）。

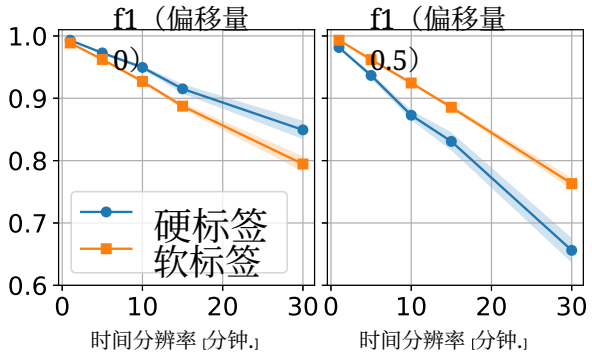

图6：不同标注时间分辨率下，硬标签和软标签相对于真实值的 F1分数。通过对真实值施加偏差并进行取整，生成标注。这里，我们将零和一半标注时间分辨率作为偏差（偏移量）。结果表明，在存在偏差时，软标签的表现优于硬标签。

不过，这留待未来工作探讨。

我们还评估了通过对真实值施加偏差（偏移量）并进行取整而生成的标注。这有助于模拟这样一种情况：标注者无法准确记住事件的起点和终点，因此除了舍入误差外，还可能引入偏差。在存在这种偏差时，预期的真实值不再与硬标签相符。我们将偏差设为标注时间分辨率的一半，并测量硬标签和软标签的 F1分数。结果（图6右侧的图）表明，在这种情况下，软标签优于硬标签。可以看出，由于标注中的偏差，硬标签的结果有所下降，而软标签的结果与之前相同。这是因为我们根据估计的标注时间分辨率，对用于生成软标签的均匀分布（式5）进行了平移。

B. 湿度事件检测

我们测试软标签和硬标签，以评估淋浴事件检测模型。该模型使用湿度传感器的读数

并预测每个时间槽是否有人在淋浴——这是一项二元分类任务。所用模型为隐马尔可夫模型。它将湿度水平作为观测值，并将淋浴状态（开/关）视为隐变量。我们假设预测模型已给定，因为本文的重点不是学习过程，而是评估。

表 I：使用软标签和硬标签的阵雨事件检测模型的混淆矩阵。

（a）使用硬标签

	预测	
标签	No	Yes
No	3186	63
Yes	41	70

（b）使用软标签

	预测	
标签	No	Yes
No	3185.14	69.50
Yes	41.86	63.50

我们使用 SRM-17 数据集 [15]，其中包含自我报告标注。我们选择了 ID:4（图 1）对应的数据集，因为其标注时间分辨率较粗。首先，我们根据自我报告标注生成硬标签和软标签，然后将其与预测模型的输出进行比较。表 I 展示了使用硬标签和软标签进行评估所得结果的混淆矩阵。我们不知道真实值，因此无法直接比较这些标签的性能。我们可以说，软标签和硬标签得出了不同的结果。关于进一步详细表征和评估的方法，我们将在本文的未来工作部分进行讨论。我们还可以看到，使用软标签所得的结果略差于使用硬标签所得的结果（真正例和真负例略少，假正例和假负例更多）。这是因为这些矩阵会对模糊性（软标签）进行惩罚；正如第 4.1 节所述，这些指标是否适用于比较性评估仍有待进一步讨论。

V. 讨论

本节首先讨论合成数据的研究发现，然后介绍其对真实世界示例的启示与局限。

类别估计的准确率： 我们的方法估计随机变量 H and C_i 的后验。该方法旨在优先选择时间分辨率较粗的标注类别，而非分辨率较细的类别。这在直觉上是合理的，因为如果标注的时间分辨率只有 30 分钟，那么自然会认为标注者采用的是 30 分钟，而不是也可能具有 0 分钟和 30 分钟刻度的更细分辨率，例如 15 分钟或 10 分钟。如果所有（或大多数）标注都落在较粗的时间分辨率上，我们的方法就会错误地估计标注类别。这种错误发生的可能性会随着

如果它有更多标注。

图7显示了标注类别估计 $\arg\max_{C_i} P(C_i|\mathbb{D})$ 的错误率。如图所示，随着接收到的标注数量增加，错误率会迅速下降。这里假设所有标注都来自单一的时间分辨率类别。每条线表示标注所对应的标注类别。没有分辨率为=30 分钟的线，因为其错误率始终为零。

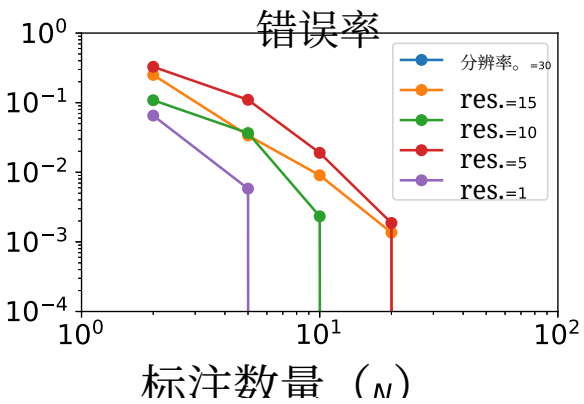

图7：标注的类别估计错误率。错误率是在合成生成的标注上测量的。每条线代表一个标注类别。X轴表示标注数量。随着标注数量增加，错误率会降低。

选择 δ 公式 *3* 中的：参数 δ 表示标注者习惯性地（H）使用不同标注类别的可能性有多大。 $\delta = 0$。1 表示标注者每十次中有一次使用不同的标注类别。它会同时影响两个随机变量后验分布 $P(H|\mathbb{D})$ 和 $P(C_i|\mathbb{D})$。如果 δ 较小（接近零），那么 H 将是与所有标注都相容的类别。例如，如果除一个标注外，其余标注的时间分辨率都是 30 分钟（例如 9:30），而其中一个标注的时间分辨率为 1 分钟（例如 13:16），那么当 δ 过小时， H 将是 c_5 （1 分钟分辨率）。另一方面，如果 δ 过大（~ 1。0），那么 C_i 的后验可能会忽略 H，而仅根据标注值 d_i选择类别。例如，如果所有标注的时间分辨率都是 5 分钟，那么 H 会正确地成为 c_4 （5 分钟分辨率）。然而，当 δ过大时，如果 $d_i \in \mathbb{C}_1$，C_i 会错误地成为 c_1。我们选择 $\delta = 0$。1，因为对于选择正确的 H 和 C_i 而言，这似乎是一个很好的平衡。

选择 $p(d_i^*|d_i, C_i)$：在本文中，我们对 $p(d_i^*|d_i, C_i)$（公式 2）采用均匀分布，因为我们假设标注者行为较为简单——仅对实际时间进行取整。然而，采用逐渐增大和减小的分布（例如高斯分布或梯形分布）会更加合理。这留待未来工作探讨。

从软标签中学习：我们提出的方法也可能应用于分类器/回归器的学习过程中。例如，通过使用所提出的软标签设置事件的起止时间，我们可以将扩增区域解释为弱标签 [14], [18], [19]。在这种情形下，可能采用的两种算法是伪标签 [20] 和乐观超集学习 [21],；它们都是迭代学习方法，会将模型在候选标签中的预测视为正确标签，并使用这些标签重新训练模型。还可以加入模型置信度较高的样本（例如置信度超过某个阈值的样本）。

我们也可以将提出的软标签视为来自伯努利分布的概率，或视为属于正类的先验信念。一些使用类似软标签的算法包括标签平滑 [22] 和期望

Jin 和 Ghahramani 提出的最大化方法 [23]。

软标注的评估指标：分类任务中常用的评估指标（例如准确率、精确率、召回率和F1分数）会因标签或预测结果中的模糊性而受到惩罚（模糊性会降低这些分数）。在本文中，我们表明，均方误差不存在此类问题，并且在实验中正确体现了软标签的优势。此外，均方误差与通常用于评估概率预测准确率的布里尔分数相似。这可能表明，均方误差是适用于软标签的正确评估指标。然而，我们仍需进一步研究，以理解在使用软标签时，哪种指标适用于哪种情境（目标）。

迈向软标签与硬标签的现实世界比较：参与者在同时进行自我标注时，对软标签和硬标签进行现实世界比较，会因缺乏完全可靠的真实值而变得复杂，尤其是在无法依赖适用于事后标注方法的数据收集的环境中（例如，视频不太适合用于家庭浴室）。然而，我们可以借助其他数据源来解决或减少部分已识别出的模糊性。例如，我们可以采用传感器融合，参考其他传感器所提供的相关信息，例如存在情况、温度或用电量，以检测家用电器的运行时间，或准确确定人员进入房间和离开房间的时间范围。这有助于解决传感器延迟这一混淆因素（例如，传感器检测到温度或湿度升高之前所需的时间），从而让我们更深入地了解现实世界中的时间安排。

时间估计中的人类表现：时间估计的表现和偏差是一个复杂的话题，其底层机制及成因大多超出了本文的讨论范围。不过，我们希望对这一机制的讨论能够激发人们进一步探索如何表征和处理自我报告数据的这一方面。

VI. 结论

我们提出了一种方法，用于识别标注者处理任务的方法及其伴随的模糊性。此外，我们设计了一种基于估计出的模糊性生成软标签的方法。评估结果表明，在均方误差方面，软标签优于硬标签。然而，就 F1分数而言，软标签的表现不如硬标签，因为 F1分数等指标本身就会惩罚模糊性。我们认为未来有许多值得探索的方向，包括改进人工标注模型、设计新的软标签评估指标，以及在学习阶段使用软标签。我们还希望在未来的论文中报告参与者对其他类型事件进行标注时粒度的方差。

致谢

本研究由英国工程与自然科学研究委员会资助项目EP/R005273/1支持的SPHERE NextSteps项目提供支持。RSR由英国研究与创新署图灵人工智能奖学金EP/V024817/1资助。

参考文献

[1] A. Möller、M. Kranz、B. Schmid、L. Roalter 和 S. Diewald，《探讨长期研究中的自我报告行为》，载于 SIGCHI 论文集，系列：CHI ’13，2013 年，第 2931–2940 页。[2] J. Lester、T. Choudhury 和 G. Borriello，《识别身体活动的实用方法》，载于国际普适计算会议。施普林格，2006 年，第 1–16 页。[3] E. L. Tonkin、A. Burrows、P. R. Woznowski、P. Laskowski、K. Y. Yordanova、N. Twomey 和 I. J. Craddock，《说、写、标？从用户视角理解智能家居数据的自我标注》，《传感器》，第 18 卷，2365，2018 年 7 月，第。[4] G. Labianca、H. Moon 和 I. Watt，《何时一小时不是 60 分钟？截止期限、时间图式以及个人和任务小组绩效》，《管理学会期刊》，第 48 卷，第 4 期，第 677–694 页，2005 年。[在线]。可访问：https://doi.org/

10 5465 2005 17843945 。/amj.。J. E. McGrath, [5]《时间在时间社会心理学中的地位》，《时间社会心理学：新视角》，第 7–17 页，1988 年。[6] F. Pan、R. Mulkar-Mehta 和 J. R. Hobbs，《为时间推理建模并学习模糊事件持续时间》，载于美国人工智能协会，2007 年，第 1659–1662 页。[7] Y. Ni、M. McVicar、R. Santos-Rodríguez 和 T. De Bie，《理解主观性对和弦估计准确率测量的影响》，《IEEE音频、语音与语言处理汇刊》，第 21 卷，第 12 期，第 2607–2615 页，2013 年。[8] S. Droit-Volet 和 W. H. Meck，《情绪如何为我们的时间感知染色》，《认知科学趋势》，第 11 卷，第 12 期，第 504–513 页，2007 年。[9] M. Wittmann 和 S. Lehnhoff，《时间感知中的年龄效应》，《心理学报告》，第 97 卷，第 3 期，第 921–935 页，2005 年。[10] S. S. Campbell、P. J. Murphy 和 C. E. Boothroyd，《长期时间估计受昼夜节律相位影响》，《生理学与行为》，第 72 卷，第 4 期，第 589–593 页，2001 年。[11] M.-C. Requena-Komuro、C. R. Marshall、R. L. Bond、L. L. Russell、C. Greaves、K. M. Moore、J. L. Agustus、E. Benhamou、H. Sivasathiaseelan、C. J. Hardy 等，《痴呆中的时间意识改变》，《神经病学前沿》，第 11 卷，第 291 页，2020 年。[12] H. Kwon、G. D. Abowd 和 T. Plötz，《处理人类活动识别中的标注不确定性》，载于第 23 届国际可穿戴计算机研讨会论文集，系列：ISWC ’19，2019 年，第 109–117 页。[13] R. Poyiadzi、W. Yang、N. Twomey 和 R. Santos-Rodríguez，《类条件标签噪声的统计假设检验》，载于欧洲机器学习与数据库知识发现原则和实践会议，2022 年。[14] R. Poyiadzi、D. Bacaicoa-Barber、J. Cid-Sueiro、M. Perello-Nieto、P. Flach 和 R. Santos-Rodriguez，《弱监督概览》，载于 2022 年 IEEE国际普适计算与通信会议研讨会及其他相关活动，2022 年，第 218–223 页。[15] T. H. Monk、E. Frank、J. M. Potts 和 D. J. Kupfer，《测量日常生活规律性的一种简单方法》，《睡眠研究期刊》，第 11 卷，第 3 期，第 183–190 页，2002 年。[16] R. A. Hauser、F. Deckers 和 P. Lehert，《帕金森病家庭日记：进一步验证及其对临床试验的启示》，《运动障碍》，第 19 卷，第 12 期，第 1409–1413 页，2004 年。[17] E. R. Coffey、E. C. Mesenbring、M. Dalaba、D. Agao、R. Alirigia、T. Begay、A. Moro、A. Oduro、Z. Brown、K. L. Dickinson 和 M. P. Hannigan，《一窥现实世界中的厨房：通过现场照片观察和改进的烹饪事件检测（cooked）分析，加深我们对炉灶使用情况的了解》，《发展工程》，第 6 卷，2021 年。[18] J. Hernández-González、I. Inza 和 J. A. Lozano，《弱监督及其他非标准分类问题：分类体系》，《模式识别快报》，第 69 卷，第 49–55 页，2016 年 1 月。[19] M. Perelló-Nieto、R. Santos-Rodríguez 和 J. Cid-Sueiro，《使监督分类算法适用于任意弱标签场景》，载于《智能数据分析进展XVI》，2017 年，第 247–259 页。[20] D.-H. Lee，《伪标签：深度神经网络的简单高效半监督学习方法》，ICML 2013研讨会：表征学习的挑战（WREPL），2013 年。[21] E. Hüllermeier 和 W. Cheng，《基于广义损失最小化的超集学习》，载于《机器学习与数据库知识发现》，2015 年，第 260–275 页。[22] R. Müller、S. Kornblith 和 G. E. Hinton，《标签平滑何时有帮助？》，载于《神经信息处理系统进展》，H. Wallach、H. Larochelle、A. Beygelzimer、F. d’Alché Buc、E. Fox 和 R. Garnett 编，第 32 卷。Curran Associates, Inc.，2019 年。[23] R. Jin 和 Z. Ghahramani，《多标签学习》，载于第 15 届神经信息处理系统国际会议论文集，系列：NIPS’02，S. Becker、S. Thrun 和 K. Obermayer 编。美国马萨诸塞州剑桥：麻省理工学院出版社，2002 年，第 921–928 页。

$$
\begin{aligned}
P(H|\mathcal{D}) &= P(\mathcal{D}|H)P(H)/P(\mathcal{D}) \\
&= \sum_{C_1,\ldots,C_N} \sum_{d_1^*,\ldots,d_N^*} P(d_1,\cdots,d_N|d_1^*,\cdots,d_N^*,C_1,\cdots,C_N)P(C_1,\cdots,C_N|H)P(d_1^*,\cdots,d_N^*)\frac{P(H)}{P(\mathcal{D})} \\
&= \sum_{C_1,\ldots,C_N} \sum_{d_1^*,\cdots,d_N^*} \prod_{i=1}^{N} P(d_i|d_i^*,C_i)P(d_i^*)P(C_i|H)\frac{P(d_1^*,\cdots,d_N^*)}{P(d_1^*)\cdots P(d_N^*)}\frac{P(H)}{P(\mathcal{D})} \\
&= \frac{P(d_1^*,\cdots,d_N^*)}{P(d_1^*)\cdots P(d_N^*)}\frac{P(H)}{P(\mathcal{D})}\prod_{i=1}^{N}\sum_{C_i} P(C_i|H)\sum_{d_i^*} P(d_i|d_i^*,C_i)P(d_i^*) \\
&= \frac{P(d_1^*,\cdots,d_N^*)}{P(d_1^*)\cdots P(d_N^*)}\frac{P(H)}{P(\mathcal{D})}\prod_{i=1}^{N}\sum_{C_i} P(C_i|H)P(d_i|C_i)
\end{aligned}
\tag{8}
$$

附录A H 后验 ($P(H|\mathbb{D})$) 推导

公式8展示了如何推导 $P(H|\mathbb{D})$。我们插入两组潜变量（C_1, $\cdots$, C_N和 d_1^*, $\cdots$, d_N^*），并应用链式法则得到第二行。然后，利用图3所示的图模型对 d_i 和C_i 进行因式分解。最后一行则通过将 $\sum$ 算子移入相关项，并将项$\sum_{d_i^*} P(d_i|d_i^*,\ C_i)d(d_i^*)$替换为 $P(d_i|C_i)$推导得出。我们还假设$P(d_1^*,\ \cdots,\ d_N^*)P(d_1^*)\cdots P(d_N^*)$= 1。~~0（假设$d^*$~~彼此相互独立），从而得到公式1。

附录B C_i 后验 ($P(C_i|\mathbb{D})$) 推导

公式9展示了 $P(C_i|\mathbb{D})$的推导。首先，我们引入隐变量 H，然后将D替换为 d_i as C_i only 取决于D中的 d_i 。接下来，我们对$P(C_i|d_i,\ H)$应用贝叶斯法则，并从 $P(d_i|C_i,\ H)$中去掉 H，因为给定 C_i 时，d_i 不取决于 H（图3）。最后，它将$P(d_i|H)$替换为 $\sum_{C_i} P(d_i|C_i)P(C_i|H)$。

$$
\begin{aligned}
P(C_i|\mathcal{D}) &= \sum_{H} P(C_i,H|\mathcal{D}) \\
&= \sum_{H} P(C_i|d_i,H)P(H|\mathcal{D}) \\
&= \sum_{H} \frac{P(d_i|C_i,H)P(C_i|H)}{P(d_i|H)}P(H|\mathcal{D}) \\
&= \sum_{H} \frac{P(d_i|C_i)P(C_i|H)}{P(d_i|H)}P(H|\mathcal{D}) \\
&= \sum_{H} \frac{P(d_i|C_i)P(C_i|H)}{\sum_{C_i} P(d_i|C_i)P(C_i|H)}P(H|\mathcal{D}),
\end{aligned}
\tag{9}
$$