Statistics and Its Interface 第0卷（2022年）1–11

错误标注数据下改进的朴素贝叶斯

Qianhan Zeng, Yingqiu Zhu*, Xuening Zhu, Feifei Wang,
Weichen Zhao, Shuning Sun, Meng Su, and Hansheng Wang

arXiv:2304.06292v1 [cs.LG] 13 Apr 2023

现实应用中经常会遇到标注错误。如果处理不当，标注错误会严重损害模型的分类性能。为解决这一问题，我们提出了一种用于文本分类的改进朴素贝叶斯方法。该方法在分析上简单明了，且无需对正确标签和错误标签作主观判断。通过指定错误标签的生成机制，我们使用 EM 算法迭代优化相应的对数似然函数。模拟和实验结果表明，对于错误标注数据，改进的朴素贝叶斯方法显著提升了朴素贝叶斯方法的性能。

AMS 2000 subject classifications: 主分类号 62F15；次分类号 62F35。
Keywords and phrases: 朴素贝叶斯，文本分类，标签噪声，EM算法。

1. 引言

近几十年来，人工智能（AI）从根本上改变了我们的工作和生活 [7, 23, 34]。人工智能算法的成功依赖于数百万甚至数十亿条标注数据 [21]。在人脸识别领域，算法经过大规模标注数据集的训练后，可以取得非凡的性能 [20]。例如，一个广受欢迎的人脸识别基准数据集是 Megaface，其中包含超过 100 万张照片。这些照片来自 690,572 名个体 [16]。在自动驾驶领域，为了训练能适应不同天气和光照条件的有效自动驾驶算法，需要大规模标注数据集。例如，nuScenes 数据集包含 140 万张图像 [5]。因此，准备大规模标注数据集是实现人工智能在现实世界中应用的重要步骤。

目前，数据标注工作主要依赖人工投入，既费力又耗时。据 Cognilytica Research 估计，AI 项目中有 [6], 25 % 的时间用于数据标注。为了节省人力，许多组织将数据标注工作外包给第三方公司，例如亚马逊 [4]。在这些公司中，标注人员应先接受培训。然而，这仍然很困难。以确保所有标注人员遵循相同的标注标准。事实上，几乎不可能定义什么是最优的标注标准。例如， [26] 发现标注人员在医学诊断中经常采用主观标准。此外，高质量的标注通常依赖大量实践。例如，[18] 发现，经验更丰富的标注人员在类别标注中的精确率和召回率通常更高。因此，在标注任务的早期阶段，标注人员更有可能出现标注错误。在某些情况下，标注任务需要专业知识，例如区分不同的鸟类物种 [25] 以及标注医学图像 [29]。错误标签可能源于标注者的系统性错误以及标注者间差异 [15, 29]。因此，在实践中，错误标签似乎是不可避免的。

事实上，许多广泛使用的数据集中都发现了错误标签。例如，[21] 报告称，QuickDraw 数据集中有 10% 的标签是错误的，亚马逊评论数据集中则有 3.9% 的标签错误。尽管实际应用中错误标签十分常见，但大多数现有研究都忽略了这一事实。它们直接使用这些数据集训练模型，将每个样本都视为标签正确。不幸的是，这种做法会降低预测性能。例如， [1] 在含有不同程度错误标签的阿拉伯手写数字数据集上训练了一个多层感知器分类器。结果表明，错误标签会显著影响分类准确率。深度神经网络模型也有类似的实证证据[32, 13]。因此，如何在开发分类方法时考虑错误标记机制，已成为一个至关重要的问题。

已有大量文献考虑了标签错误问题。这些研究大致可分为两类，分别是噪声过滤方法和改进的模型架构。噪声过滤方法的思路是，首先通过机器学习方法区分标注正确的实例和标注错误的实例，然后仅基于标注正确的实例构建模型[3, 27, 12, 8, 14]。在这方面，如何区分标注错误的实例与标注正确的实例至关重要。这引出了两种可能的解决方案。第一种是将预测标签与观测标签不同的实例视为标注错误的

*通讯作者。

实例。然而，这种方法可能存在问题，因为即使对于标注正确的实例，其观测标签和预测标签也可能不同。这种差异并非由标注错误造成。第二种解决方案是通过人工投入生成一个完全正确标注的数据集。然而，这在人力成本上可能非常昂贵，也会面临标注者间差异。

除了噪声过滤方法外，许多研究者还通过修改模型架构来处理错误标记问题。BayesANIL[24] 提出了一种贝叶斯模型，用于处理近似、含噪或不完整标签的学习。然而，他们的方法需要预先指定噪声率，而在实践中这一数值很难获知。其他典型方法着眼于修改深度学习模型，例如含噪标签神经网络（NLNN）算法 [2], 解耦方法 [19], 和协同教学方法[11]。然而，这些经过修改的深度学习模型通常包含大量未知参数。例如，NLNN 算法包含超过 50 万个参数，而解耦方法和协同教学方法均包含超过 800 万个参数。

为解决上述问题，我们在此提出一种改进的朴素贝叶斯模型，用于处理含有错误标签的文档分类。我们的算法源自朴素贝叶斯模型，该模型已被证实可谓是文档分类中最有用的工具之一[28, 8]。为处理错误标记问题，我们使用错误标记概率矩阵对错误标记机制进行建模。据此，可以明确写出对数似然函数。这使我们有机会研究可识别性问题，并开发用于模型估计的期望最大化算法。在真实世界基准数据集上的模拟和实验表明，我们的算法优于传统朴素贝叶斯模型及其他一些基线方法。

2. 相关工作

如前所述，使用错误标注数据进行分类是一个至关重要的问题。现有文献采用了不同策略来解决这一问题。这些方法大致可分为两类。第一类使用噪声过滤方法，为后续数据分析任务生成干净数据。第二类则侧重于修改模型架构，使模型能够考虑错误标注行为。下面我们分别总结这两类相关工作。

噪声过滤方法。处理错误标注实例的一种方法是使用噪声过滤方法。这类方法的基本思想是训练模型，以区分标注正确的实例和错误标注的数据。随后，仅在标签正确的样本上执行后续数据分析任务。例如，[3] 应用了一系列学习算法（如决策树、 k近邻分类器和线性分类器）来构建噪声过滤器。预测标签与观测标签不同的样本会被作为错误标注数据丢弃。[27] 在训练数据上构建了支持向量机作为过滤器，并移除了所有被该过滤器错误分类的样本。 [12]针对二分类问题构建了一对名为真值神经网络（Truth NN）和假值神经网络（False NN）的神经网络，因此每个标签只能是 0 或 1。这两个网络具有相同的结构和输入。不过，真值神经网络使用观测标签进行训练。相反，假值神经网络使用观测标签的补数进行训练。二分类标签的补数定义为 1 减去该标签。被真值神经网络和假值神经网络同时错误预测的样本被视为错误标注样本。 [8] 采用朴素贝叶斯分类器，根据最高后验概率为每个训练样本预测类别标签。随后，将标签与预测标签不同的训练样本从训练数据集中移除，以便为后续分析生成无噪声数据。 [14]首先需要一个所有标签都正确的小型数据集，以学习数据驱动课程。随后，使用该数据集训练一个名为 MentorNet 的深度学习模型，为错误标注数据提供初步信息。最后，构建名为 StudentNet 的基础深度卷积神经网络（CNN），并在训练过程中与 MentorNet 一同更新。

噪声过滤方法的优点在于简单直观，但也面临一些挑战。具体而言，要应用噪声过滤方法，必须判断某个样本是否被正确标注。在这方面，通常采用两种策略。一种是比较观测标签与预测标签，并在两者不一致时将样本视为错误标注。这种策略非常主观，因为即使所有观测标签都正确，大多数分类器仍可能产生错误预测。因此，出现不一致时，很难判断该样本究竟是错误标注，还是仅仅预测错误。第二种策略是通过人工投入生成完全正确标注的数据集，但这会带来额外的标注成本。此外，人工标注过程也可能产生错误标签。换言之，在实践中，创建完全正确的数据集既昂贵又困难。

修改模型架构的方法。处理错误标记样本的另一种方法是直接修改模型架构。这类方法的基本思路是在概率框架或非参数框架下对错误标记行为进行建模。例如，

[24] 构建了一个名为 BayesANIL 的贝叶斯模型，并将真实标签视为潜在变量。随后，他们应用期望最大化算法来处理未标记和错误标记的样本。 [2] 提出了含噪标签神经网络（NLNN）模型，该模型假设观测标签是由真实标签经过一个噪声通道生成的。随后，他们在神经网络中引入了一个额外的噪声层，以对含噪标签进行建模。 [10] 在 [2] 的工作基础上进行了扩展，允许含噪标签的出现不仅取决于真实标签，也取决于输入特征。 [19] 通过训练两个网络提出了解耦方法。在每次小批量迭代中，两个网络预测标签不一致的样本会被挑选出来。然后，两个网络利用这些挑选出的样本更新参数。[11] 通过训练两个神经网络提出了协同教学方法。每个网络在每个小批量上进行前向传播，并选出损失较小的数据作为干净数据。随后，每个网络都利用由其同伴网络提供的、经过筛选的干净数据进行反向传播。

这些方法的优点在于，在模型输入阶段无需将错误标记的实例与正确标记的实例区分开来。然而，这些方法仍然存在一些问题。首先，有些模型通常非常复杂。例如，NLNN 模型包含 50 万个参数 [2, 10]。解耦方法和协同教学方法需要更新超过 800 万个参数 [19, 11]。其次，这一策略背后存在一个强假设，即所有错误标签都应具有相同的模式，这样才能建立模型来发现这一模式并识别出错误标签。然而，在某些情况下，错误标签的产生完全是随机的。因此，模型几乎不可能准确预测这些随机生成的错误标签。最后，一些方法需要预先了解某些重要的调优参数。例如，应用 BayesANIL 方法需要了解噪声率 [24], , 而这一参数在实践中通常是未知的。

我们的贡献。受现有文献启发，我们为文本分类开发了一种改进的 Na¨ıve Bayes 模型。我们关注文本分类，因为它具有广泛的应用。改进后的 Na¨ıve Bayes 模型在分析上简洁明了，且无需对正确标签和错误标签作主观判断。通过假设真实标签不可观测，并明确错误标签的生成机制，可以构造相应的对数似然函数，并开发期望最大化算法。在真实世界基准数据集上的模拟和实验表明，我们的算法优于传统的 Na¨ıve Bayes 分类器及其他一些基线方法。

¨3. 含错误标注数据的改进朴素贝叶斯

3.1 符号与模型

假设共有 N 个实例。每个实例属于 K 个类别之一。设 $Y_i^* \in \{1, \cdots, K\}$ 为第 i个实例的真实标签，其中 $1 \le i \le N$。对于每个实例，我们收集一个 d维二元特征向量。例如，对于文档分类问题，每篇文档都视为一个实例。文档的一个特征可以定义为该文档是否包含某个特定词。若该词出现在文档中，则该特征的值为 1，否则为 0。对于第 i个实例，其二元特征向量记为 $X_i = (X_{i1}, ..., X_{id})^\top$。我们用 $\mathbb{X} = \{X_i, 1 \le i \le N\}$表示所有实例。所有实例的真实标签记为 $\mathbb{Y}^* = \{Y_i^*, 1 \le i \le N\}$。

为了研究训练样本 $\mathbb{X}$ 与真实标签 $\mathbb{Y}^*$之间的关系，可以通过最大化对数似然函数 $\ln P(\mathbb{X}, \mathbb{Y}^*)$ 来应用典型的朴素贝叶斯模型。令 $\pi_k = P(Y_i^* = k)$ 表示类别 k的概率。于是有 $\sum_{k=1}^K \pi_k = 1$。令 $p_{jk} = P(X_{ij} = 1 | Y_i^* = k)$ 表示类别 k中第 j个特征取值为 1 的概率，并令 θ 为总参数集，即 $\theta = \{p_{jk}, 1 \le j \le d, 1 \le k \le K\} \cup \{\pi_k, 1 \le k \le K\}$。为了估计 θ，我们可以构造如下对数似然函数

$$
\begin{aligned}
\ell(\theta) = \ln P(\mathbb{X}, \mathbb{Y}^*) &= \sum_{i=1}^N \ln P(X_i, Y_i^*) \\
&= \sum_{i=1}^N \ln P(Y_i^*) + \sum_{i=1}^N \sum_{j=1}^d \ln P(X_{ij} | Y_i^*) \\
&= \sum_{i=1}^N \ln \pi_{Y_i^*} + \sum_{i=1}^N \sum_{j=1}^d X_{ij} \ln p_{jY_i^*} \\
&\quad + \sum_{i=1}^N \sum_{j=1}^d (1 - X_{ij}) \ln \left(1 - p_{jY_i^*}\right).
\end{aligned}
$$

Na¨ıve Bayes 模型的目标是最大化对数似然函数 $\ell(\theta)$。然而，只有在所有真实标签都可观测的情况下，才能进行上述估计。尽管如此，现实中由于错误标记问题，观测标签可能并不正确，这种情况十分常见。简单地将观测标签视为真实标签，可能会降低 Na¨ıve Bayes 模型的估计性能。我们将在后续章节讨论如何解决错误标记问题。

3.2 错误标记机制

首先，我们指定错误标记机制，以便进行后续的模型估计。我们假设真实标签 Y_i^*是一个潜在变量，观测到的标签记为 Y_i。接着，我们将所有实例的观测标签记为

$\mathbb{Y} = \{Y_i, 1 \le i \le N\}$。遵循文献[22, 33],，我们假设 $P(Y_i|Y_i^*, X_i) = P(Y_i|Y_i^*)$。这样，我们假设误标模式是随机的，在给定真实标签 X_i 的条件下，不依赖于特征 Y_i^*。接下来，我们考虑如何恰当地建立错误标记机制$P(Y_i \neq Y_i^*|Y_i^*$。

定义 $\rho_{k_1k_2}$ 为标注者在真实类别为 k_2 时，将一个实例标记为类别 k_1 的概率，其中 $1 \le k_1, k_2 \le K$。由此得到如下错误标记概率矩阵

$$\begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1K} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{K1} & \rho_{K2} & \cdots & \rho_{KK} \end{pmatrix}, \tag{1}$$

其中 $\sum_{k_1=1}^{K} \rho_{k_1k_2} = 1$ 对于 $1 \le k_2 \le K$。可以看出，若 $\rho_{kk} = 1$ 对于 $1 \le k \le K$，则该模型退化为标准的朴素贝叶斯模型。否则，它将成为一个新模型。若对于任意 $\rho_{k_1k_2} = \rho_{k_2k_1}$ $k_1 \neq k_2$，我们便将（1）称为对称误标矩阵。在此模型设定下，整个参数集变为 $\theta = \{p_{jk}, 1 \le j \le d, 1 \le k \le K\} \cup \{\pi_k, 1 \le k \le K\} \cup \{\rho_{k_1k_2}, 1 \le k_1, k_2 \le K\}$。相应的对数似然函数为

$$\begin{aligned} \ell(\theta) &= \ln P(\mathbb{X}, \mathbb{Y}, \mathbb{Y}^*|\theta) \\ &= \sum_{i=1}^{N} \left\{ \ln P(Y_i^*|\theta) + \ln P(Y_i|Y_i^*, \theta) + \sum_{j=1}^{d} \ln P(X_{ij}|Y_i^*, \theta) \right\} \\ &= \sum_{i=1}^{N} \ln \pi_{Y_i^*} + \sum_{i=1}^{N} \ln \rho_{Y_iY_i^*} + \sum_{i=1}^{N}\sum_{j=1}^{d} X_{ij} \ln p_{jY_i^*} \\ &\quad + \sum_{i=1}^{N}\sum_{j=1}^{d} (1 - X_{ij}) \ln (1 - p_{jY_i^*}). \end{aligned} \tag{2}$$

乍看之下，我们似乎无法排除这样一种理论上的可能性：$\rho_{kk} < \max_{k'} \rho_{k'k}$ 对于 $1 \le k \le K$在这种错误标记机制下。如果出现这种情况，可能存在一个 $k \in \{1, \cdots, K\}/\{Y_i^*\}$，使得 $P(Y_i = k|Y_i^*) > P(Y_i = Y_i^*|Y_i^*)$。这意味着，可能存在一个错误类别，其对应的错误标记概率甚至高于正确标注的概率。这似乎非常不合理。我们进一步的理论分析表明，如果不对式（1）中的错误标记概率矩阵施加任何约束，错误标记机制可能会面临可识别性问题。我们将在下一小节介绍这一可识别性问题。

3.3 可识别性问题

如前所述，错误标记机制可能存在可识别性问题。其主要原因如下。注意，真实标签 $\mathbb{Y}^*$ 是潜在的。若对错误标记概率矩阵不施加任何约束，那么只要能够复现观测标签 $\mathbb{Y}$ 的概率分布，任意指定 $\mathbb{Y}^*$ 都是可以接受的。例如，考虑一个真实标签为 $Y_i^* \in \{1, \cdots, K\}$、观测标签为$Y_i \in \{1, \cdots, K\}$的分类问题。这里，我们将真实标签 Y_i^*对 $Y_i^* \in \{1, \cdots, K-1\}$ 平移至 $Y_i^* + 1$，并将 $Y_i^* = K$ 替换为 1。我们~将平移后的标签记为 Y_i^*，并将所有平移后的~标签汇总为 $\mathbb{Y}^*$。将从 Y_i^*到~Y_i^*的一一映射记为 $g(\cdot)$。回顾一下，基于$\mathbb{Y}^*$ 的对数似然函数如式（2）所示。然而，将 $\rho_{k_1k_2}$和 p_{jk}平移后，基于 $\mathbb{Y}^*$的新对数似然函数值~保持不变。新的标记概率 $\widetilde{\rho}_{k_1k_2}$ 满足 $\widetilde{\rho}_{k_1k_2} = \rho_{k_1g^{-1}(k_2)}$，新的特征概率 $\widetilde{p}_{jk}$~满足 $\widetilde{p}_{jk} = p_{jg^{-1}(k)}$。将新的参数集记为 θ。~因此，基于 $\mathbb{Y}^*$的新对数似然函数变为

$$\begin{aligned} \widetilde{\ell}\left(\widetilde{\theta}\right) &= \ln \left(\mathbb{X}, \mathbb{Y}, \widetilde{\mathbb{Y}}^* \middle| \widetilde{\theta}\right) \\ &= \sum_{i=1}^{N} \left\{ \ln P\left(\widetilde{Y}_i^* \middle| \widetilde{\theta}\right) + \ln P\left(Y_i \middle| \widetilde{Y}_i^*, \widetilde{\theta}\right) + \sum_{j=1}^{d} \ln P\left(X_{ij} \middle| \widetilde{Y}_i^*, \widetilde{\theta}\right) \right\} \\ &= \sum_{i=1}^{N} \ln \pi_{\widetilde{Y}_i^*} + \sum_{i=1}^{N} \ln \widetilde{\rho}_{Y_i\widetilde{Y}_i^*} + \sum_{i=1}^{N}\sum_{j=1}^{d} X_{ij} \ln \widetilde{p}_{j\widetilde{Y}_i^*} \\ &\quad + \sum_{i=1}^{N}\sum_{j=1}^{d} (1 - X_{ij}) \ln (1 - \widetilde{p}_{j\widetilde{Y}_i^*}). \end{aligned}$$

很容易验证，得到的对数似然值~~与之前完全相同，即 $\ell(\theta) = \ell(\theta)$。这表明，关于~ $\mathbb{Y}$和 $\mathbb{X}$ 的观测概率行为，可以用不同参数下的 $\mathbb{Y}^*$ 或~ $\mathbb{Y}^*$ 同样充分地解释。因此，基于这种错误标记机制的模型可能无法被唯一识别。那么，如何解决这一可识别性问题便成为一个重要问题。对此，我们注意到，合格的标注者不应表现得太差，也就是说，对角线概率应大于非对角线概率。否则，可能存在一个类别 $k_1 \neq k_2$满足 $P(Y_i = k_1|Y_i^* = k_2) > P(Y_i = k_2|Y_i^* = k_2)$。因此，我们假设对角线元素 ρ_{kk}大于非对角线元素 $\rho_{k'k}$ 且 $k' \neq k$，即对每个 $1 \le k \le K$都有$\rho_{kk} > \max_{k' \neq k} \rho_{k'k}$。这使得基于（1）的模型具有可识别性。

3.4 错误标签的影响

正如我们之前所展示的，错误标签可能会，也可能不会严重影响模型的性能 [1]。在本小节中，我们从理论上研究误标的影响，并说明误标何时对预测影响甚微，何时会造成严重问题。我们先从一个带有 $K = 2$ 的简单案例开始。

二分类情形。当只有两个类别时，式（1）中的错误标记概率矩阵变为 2×2 矩阵。为简单起见，这里假设每个类别的样本量均衡，且 $P(Y_i^* = 1) = P(Y_i^* = 2) = 0.5$。为便于说明，我们进一步假设 $\rho_{12} = \rho_{21}$ 因此

使得 $P(Y_i = 1) = P(Y_i = 2) = 0.5$。回想一下，在类别 k 中，第 j个特征取值为 1 的概率记为p_{jk}，对于任意 $k \in \{1, 2\}$， $P(X_{ij} = 1|Y_i^* = k)$。为便于直观理解，不失一般性地假设 $p_{j1} > p_{j2}$。这表明，随机事件 $X_{ij} = 1$发生的可能性在 $Y_i^* = 1$ 下比在 $Y_i^* = 2$下更高。这进一步意味着，若可以假设 Y_i^*的先验概率相等，则 $P(Y_i^* = 1|X_{ij} = 1) > P(Y_i^* = 2|X_{ij} = 1)$。因此，如果我们掌握的唯一信息是 $X_{ij} = 1$ ，那么自然应预测 Y_i^*取值为 1。接下来，一个很值得探讨的问题是，错误标记会如何影响这一点。我们尤其关心 $P(Y_i = 1|X_{ij} = 1)$ 和 $P(Y_i = 2|X_{ij} = 1)$ 的相对大小关系。于是有

$$\begin{aligned} &P(Y_i = 1|X_{ij} = 1) - P(Y_i = 2|X_{ij} = 1) \\ &= 0.5 \times (p_{j1} - p_{j2}) \times (\rho_{11} - \rho_{12})/P(X_{ij} = 1) > 0, \end{aligned} \tag{3}$$

由于 $p_{j1} - p_{j2} > 0$ 且 $\rho_{11} - \rho_{12} > 0$ 根据模型假设。根据(3)，我们发现 $P(Y_i = 1|X_{ij} = 1)$ 与 $P(Y_i = 2|X_{ij} = 1)$ 的相对顺序不会因错误标记而改变。因此，如果 $X_{ij} = 1$ 是我们掌握的唯一信息，我们自然应预测 Y_i 为 1。该预测结果与 Y_i^*相同。这表明，对于这一特定的特殊情况，错误标记造成的不利影响似乎非常有限。

具有恒定错误标记概率的多类别情况。上述讨论表明，如果 $K = 2$ 且错误标记概率矩阵是对称的，那么错误标记造成的不利影响可能非常小。因此，我们受到启发，研究 $K \geq 3$ 的情况。为简单起见，我们考虑错误标记概率恒定的情形。也就是说，假设对于某个 $\rho \in (0.5, 1)$， $\rho_{kk} = \rho$ ，且对于任意 $k_1 \neq k_2$， $\rho_{k_1 k_2} = (1 - \rho)/(K - 1)$。这里我们假设每个类别的先验概率相等。不失一般性，我们可以假设 $p_{jk_1} > p_{jk_2}$。这意味着随机事件 $X_{ij} = 1$ 发生的可能性高于 $Y_i^* = k_2$，即 $Y_i^* = k_1$。因此，如果观察到 $X_{ij} = 1$ ，那么预测 $Y_i^* = k_1$ 比预测 $Y_i^* = k_2$ 更合理。事实上，可以验证，若假设 Y_i^* 的先验概率相等，则 $P(Y_i^* = k_1|X_{ij} = 1) > P(Y_i^* = k_2|X_{ij} = 1)$。具体而言，我们还可以通过以下方式计算 $P(Y_i = k_1|X_{ij} > 1) - P(Y_i = k_2|X_{ij} > 1)$】【。

$$\begin{aligned} &P(Y_i = k_1|X_{ij} > 1) - P(Y_i = k_2|X_{ij} > 1) \\ &= \left(\frac{1}{K}\right) \times (p_{jk_1} - p_{jk_2}) \times \left(\frac{K\rho - 1}{K - 1}\right) / P(X_{ij} = 1) > 0, \end{aligned} \tag{4}$$

与 $K \geq 3$ 和 $\rho > 0.5$。根据（4），对于任意两个类别 $P(Y_i = k_1|X_{ij} = 1)$ 与 $P(Y_i = k_2|X_{ij} = 1)$，它们的相对顺序与 $P(Y_i^* = k_1|X_{ij} = 1)$ 和 $P(Y_i^* = k_2|X_{ij} = 1)$ 的相对顺序保持一致。因此，基于观测标签的朴素贝叶斯模型，其预测结果与基于真实标签的模型相同。这表明，只要错误标记概率保持恒定，错误标记造成的不利影响就非常有限。

错误标记概率变化的多类别情形。接下来，我们研究具有 $K \geq 3$ 和非恒定错误标记概率的情形。具体而言，我们在此考虑一种特殊的 $K \geq 3$ 情形。错误标记概率矩阵如下所示

$$\begin{pmatrix} \rho & 1-\rho & 1-\rho & \cdots & 1-\rho \\ 1-\rho & \rho & 0 & \cdots & 0 \\ 0 & 0 & \rho & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \rho \end{pmatrix}, \tag{5}$$

其中 $0.9 \leq \rho < 1$。在此情况下，我们要求 $K \gg \rho/(1-\rho) + 1 \geq 11$。根据这个错误标记概率矩阵，我们假设类别 1 最容易与其他类别混淆。具体而言，我们知道，对于任意 $k_1 > 2$， $k_2 > 2$， $\rho_{k_1 k_2} = 0$ ，且 $k_1 \neq k_2$。换句话说，除非某个类别是类别 1，否则任意类别对之间都不会发生错误标记。如果真实类别是 $Y_i^* = 1$，那么以 ρ 的概率，观测类别 Y_i 仍为 1；以 $(1 - \rho)$ 的概率，观测类别 Y_i 会变为 2。如果真实类别 Y_i^* 是类别 $k \geq 2$，那么以 ρ 的概率，观测类别 Y_i 仍为 k ；以 $(1 - \rho)$ 的概率，观测类别 Y_i 会变为 1。为简单起见，我们假设对于 $k \geq 2$， $p_{j1} \ll p_{jk}$ 。这导致 $P(Y_i^* = 1|X_{ij} = 1) < P(Y_i^* = k|X_{ij} = 1)$。然而，很容易验证

$$\begin{aligned} &P(Y_i = 1|X_{ij} = 1) - P(Y_i = k|X_{ij} = 1) \\ &= \left(\frac{1-\rho}{K}\right) \times \left\{ \left(\frac{\rho}{1-\rho}\right) p_{j1} + (K-1)p_{j2} - \left(\frac{\rho}{1-\rho}\right) p_{j2} \right\} > 0, \end{aligned}$$

对于任意 $k > 2$，情况都是如此。因此，错误标记机制会造成严重的不利影响。 consequently，朴素贝叶斯模型几乎无法区分类别 1 和类别 $k > 2$。这一特殊情况表明，当错误集中累积在少数几个类别上时，朴素贝叶斯模型可能难以正确识别这些类别。这种情况下，即使只有很小一部分错误标签，也仍然会造成严重问题，而无法通过简单地扩大样本量来解决。

请注意，标注错误可能会严重降低分类准确率。因此，如何评估错误标记效应就成为一个非常重要的问题。请注意，分类准确率是分类性能评估的一项重要指标。将基于 $(\mathbb{X}, \mathbb{Y})$ 的分类准确率记为 ACC。将基于 $(\mathbb{X}, \mathbb{Y}^*)$ 的分类准确率记为 ACC^*。我们可以将它们的差值 $\Delta\text{ACC} = \text{ACC} - \text{ACC}^*$ 定义为衡量错误标记效应的一种简单指标。

3.5 一种期望最大化算法

如我们在小节 3.3中所示，通过假设错误标记机制中的错误标记概率矩阵 (1) 满足 $\rho_{kk} > \max_{k' \neq k} \rho_{k'k}$ 对每个 $1 \leq k \leq K$，由此推导出的对数似然函数是可识别的。具体而言，我们考虑如何估计

通过优化对数似然函数 $\ell(\theta)$ 来估计相应模型。由于真实标签 $\mathbb{Y}^*$ 未知，无法直接优化 $\ell(\theta)$。因此，我们提出一种期望最大化算法，通过迭代估计参数并优化对数似然函数。每次迭代中，都会执行期望步骤（E步）和最大化步骤（M步）。下面将详细讨论每个步骤。

E步。在第 t次迭代中，我们计算给定观测数据 $(\mathbb{X}, \mathbb{Y})$ 和当前参数估计 $\widehat{\theta}^{(t)}$时，对数似然 $\ell(\theta)$的期望。条件期望可以推导为

$$
\begin{aligned}
Q\left(\theta, \widehat{\theta}^{(t)}\right) =& E\left\{\ell(\theta)\middle|\mathbb{X}, \mathbb{Y}, \widehat{\theta}^{(t)}\right\} \\
=& \sum_{i=1}^{N}\sum_{k=1}^{K}\widehat{\gamma}_{ik}^{(t)} \ln \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K}\widehat{\gamma}_{ik}^{(t)} \ln \rho_{Y_i k} \\
&+ \sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{k=1}^{K} X_{ij}\widehat{\gamma}_{ik}^{(t)} \ln p_{jk} \\
&+ \sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{k=1}^{K}(1-X_{ij})\widehat{\gamma}_{ik}^{(t)} \ln (1-p_{jk}),
\end{aligned} \tag{6}
$$

其中 $\widehat{\gamma}_{ik}^{(t)}$表示 $P\left(Y_i^* = k | X_i,\ Y_i,\ \widehat{\theta}^{(t)}\right)$。式（6）的计算取决于 $\widehat{\gamma}_{ik}^{(t)}$，其中 $1 \le i \le N$ 以及 $1 \le k \le K$。为计算 $\widehat{\gamma}_{ik}^{(t)}$，我们首先定义 $\widehat{\zeta}_{ik}^{(t)}$ 如下

$$
\widehat{\zeta}_{ik}^{(t)} = \widehat{\pi}_k^{(t)} \widehat{\rho}_{Y_i k}^{(t)} \prod_{j=1}^{d} \widehat{p}_{jk}^{(t) X_{ij}} \left\{1 - \widehat{p}_{jk}^{(t)}\right\}^{1-X_{ij}}. \tag{7}
$$

然后我们计算 $\widehat{\gamma}_{ik}^{(t)} = \widehat{\zeta}_{ik}^{(t)} / \sum_{k=1}^{K} \widehat{\zeta}_{ik}^{(t)}$。

M步。在此步骤中，我们最大化条件期望 $Q(\theta, \widehat{\theta}^{(t)})$，以得到新的估计值 $\widehat{\theta}^{(t+1)}$。注意，对于 $1 \le k_2 \le K$，$\sum_{k=1}^{K} \pi_k = 1$ 和 $\sum_{k_1=1}^{K} \rho_{k_1 k_2} = 1$。根据拉格朗日乘数法，我们可以将拉格朗日函数定义为

$$
Q = Q\left(\theta, \widehat{\theta}^{(t)}\right) + \lambda\left(1 - \sum_{k=1}^{K}\pi_k\right) + \sum_{k_2=1}^{K}\lambda_{k_2}\left(1 - \sum_{k_1=1}\rho_{k_1 k_2}\right),
$$

其中，λ 和 λ_{k_2} 以及 $1 \le k_2 \le K$ 是拉格朗日乘数法中使用的附加参数。最大化拉格朗日函数 Q 后，第 $(t+1)$ 次迭代中的更新参数计算如下：

$$
\begin{aligned}
\widehat{\pi}_k^{(t+1)} &= \sum_{i=1}^{N}\widehat{\gamma}_{ik}^{(t)}/N,\ 1 \le k \le K, \\
\widehat{p}_{jk}^{(t+1)} &= \left(\sum_{i=1}^{N} X_{ij}\widehat{\gamma}_{ik}^{(t)}\right) \Big/ \left(\sum_{i=1}^{N}\widehat{\gamma}_{ik}^{(t)}\right),\ 1 \le j \le d,\ 1 \le k \le K, \\
\widehat{\rho}_{k_1 k_2}^{(t+1)} &= \left(\sum_{i=1}^{N} I(Y_i = k_1)\widehat{\gamma}_{ik_2}^{(t)}\right) \Big/ \left(\sum_{i=1}^{N}\widehat{\gamma}_{ik_2}^{(t)}\right),\ 1 \le k_1, k_2 \le K.
\end{aligned}
$$

为了初始化参数，我们可以为 $1 \le k \le K$ 设置 $\widehat{\pi}_k = 1/K$。此外，对于 $\widehat{p}_{jk}$ $(1 \le j \le d, 1 \le k \le K)$，我们采用随机初始化。对于 $1 \le k_1, k_2 \le K$，我们对 $\widehat{\rho}_{k_1 k_2}$进行随机初始化，并使对角元素大于 0.5。随后，反复执行 E步和 M步，直到收敛或达到预先设定的最大迭代次数。

4. 实验

4.1 模拟实验

我们进行了数值研究，以评估所提方法的有限样本性能。

模拟数据。我们生成合成数据，并将实例的维度设为 $d = 500$。每个实例属于 $K = 5$ 个类别之一。数据规模设为 $n = 500$、1000 和 5000。将 20% 的模拟数据划分为测试集。先验概率 π_k 设为 $1/K$，适用于任意 $1 \le k \le K$。附录 A.2中包含一个不平衡情形。错误标记概率矩阵的对角元素从不同区间中随机生成，如表 1所示。随后从 $[0, 1)$ 中随机抽取非对角元素，使得 $\sum_{k_1=1}^{K} \rho_{k_1 k_2} = 1$ 对任意$1 \le k_2 \le K$均成立。通过将 p_{jk} 与从 $[0, 0.1)$ 中生成的均匀随机数，以及均值为 0.65、标准差为 0.06 的正态随机数相加，生成这些值。在每种模拟设置下，实验重复 $B = 100$ 次。

基线方法。我们将改进的朴素贝叶斯方法（INB）与以下方法进行比较：1）不考虑错误标记的朴素贝叶斯（NB）模型；2） [2]的NLNN方法；3） [10]的NAL[1] 方法；4）使用参数真实值的NB模型（NB-T）。

评估指标。我们从两个方面评估所提出的方法。首先，使用均方误差（MSE）评估参数估计的准确性。MSE定义为估计参数与参数真实值之差的平方的均值。MSE $= \sum_{j=1}^{d}\sum_{k=1}^{K}(\widehat{p}_{jk} - p_{jk})^2/(dK)$，用于衡量估计准确性。其次，使用测试集上的分类准确率（ACC）评估分类准确率，即 ACC $= s/n_0 \times 100\%$，其中 s 是测试集中正确预测的标签数，n_0 是测试集的大小。第三，我们还计算了各方法的曲线下面积（AUC）[2]（即受试者工作特征曲线，ROC曲线，下方的面积） [31]。这里使用的AUC是宏平均AUC指标，它是一种多类别性能度量。第四，我们还评估了模拟数据上的错误标记效应ΔACC，其定义见第3.4节。

[1] https://github.com/udibr/noisy labels [2] https://scikit-learn.org/stable/modules/model evaluation.html#roc- metrics

表1. 不同 ρ_{kk} 区间和样本量下不同方法的有限样本性能。

ρ_{kk} 区间	n	均方误差×10^{-3} NB	均方误差×10^{-3} INB	准确率（%） NB	准确率（%） INB	准确率（%） NB-T	准确率（%） NLNN	准确率（%） NAL	曲线下面积（%） NB	曲线下面积（%） INB	曲线下面积（%） NB-T	曲线下面积（%） NLNN	曲线下面积（%） NAL	Δ准确率（%）
[0.55, 0.65)	500	3.3	2.9	66.0	83.2	96.6	20.7	28.3	90.7	97.2	99.7	52.7	63.7	-22.3
	1000	2.2	1.2	75.9	92.6	96.9	21.1	39.3	94.7	99.4	99.8	55.4	74.8	-17.6
	5000	1.0	0.2	90.9	95.0	95.7	30.8	80.6	99.1	99.7	99.8	64.4	96.5	-4.2
[0.65, 0.75)	500	3.0	2.8	73.8	84.0	96.6	20.7	30.5	94.2	97.7	99.7	53.8	67.8	-14.5
	1000	1.7	1.2	85.3	92.7	96.9	21.1	46.8	97.9	99.4	99.8	56.4	80.6	-8.1
	5000	0.7	0.2	93.3	95.1	95.7	35.3	86.1	99.5	99.7	99.8	66.7	98.0	-1.8
[0.75, 0.85)	500	2.7	2.7	78.7	85.6	96.6	20.8	32.9	96.0	98.1	99.7	55.0	70.0	-9.6
	1000	1.4	1.2	89.2	93.0	96.9	22.0	54.6	98.8	99.4	99.8	58.6	85.2	-4.2
	5000	0.4	0.2	94.2	95.1	95.7	34.9	88.6	99.6	99.7	99.8	65.6	98.6	-0.9
[0.85, 0.95)	500	2.4	2.5	84.3	86.8	96.6	21.7	36.8	97.9	98.4	99.7	56.7	73.3	-4.0
	1000	1.2	1.2	91.8	93.2	96.9	22.3	60.0	99.3	99.5	99.8	58.7	88.1	-1.6
	5000	0.3	0.2	94.7	95.1	95.7	35.6	90.1	99.7	99.7	99.8	65.2	99.0	-0.4
[1.0, 1.0]	500	2.3	2.4	88.2	88.0	96.6	22.1	40.0	98.8	98.7	99.7	56.9	76.2	0.0
	1000	1.1	1.1	93.4	93.3	96.9	24.0	64.5	99.5	99.5	99.8	61.7	90.4	0.0
	5000	0.2	0.2	95.1	95.1	95.7	31.2	92.1	99.7	99.7	99.8	62.4	99.3	0.0

含 $0.5 < \rho_{kk} < 1$ 的模拟结果。根据表 Δ1最后一列的准确率，可知随着 ρ_{kk}s 增大，错误标记效应 ΔACC 会降低。这是合理的，因为随着 ρ_{kk}s 增大，错误标签的数量会减少，因此错误标记效应得到缓解。不同 ρ_{kk} 区间和数据规模下的MSE、ACC和AUC结果见表1。图 [9] 1展示了ROC曲线示例，该曲线是不同类别ROC曲线的平均值。首先可以看到，随着数据规模增大，不同方法的性能均有所提升。其次，随着 ρ_{kk}s 增大，不同方法的平均性能有所提升，且彼此更加接近。由于实际应用中参数的真实值未知，NB-T方法不适用于实际应用。在模拟研究中，NB-T方法的性能可作为理论最优分类性能的重要参照。与NB方法相比，INB方法的性能更接近NB-T方法。最后，除NB-T方法外，在大多数情况下，INB方法在所有模拟设置下均优于其他方法。具体而言，在不同 ρ_{kk} 区间和数据规模下，INB方法在测试集上大多数情况下取得了最小的MSE值、最高的ACC值和最高的AUC值。这体现了所提出的INB方法的优势。

$\rho_{kk} = 1$的模拟结果。请注意，如果对于任何 $1 \le k \le K$， $\rho_{kk} = 1$，则模拟实验中不会生成错误标签。因此，错误标记效应Δ准确率自然为 0。当没有发生错误标记时，NB、INB 和 NB-T 的性能相同且最优。

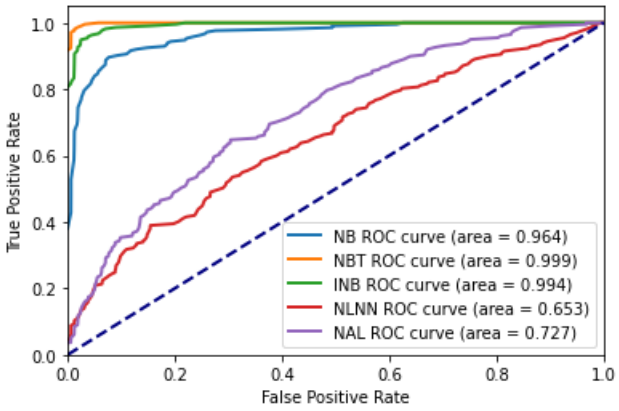

图1. 任意选取的 ROC曲线示例，包含 $n = 1000$ 和 $\rho_{kk} \in [0.55, 0.65)$ 。

4.2 真实数据实验

20新闻组基准数据集。我们在一个重要的基准数据集上比较了各方法的性能20新闻组数据集 [17]。该数据集包含18,846篇文档，其中训练集有15,076篇，测试集有3,770篇。首先，我们为这些文档构建了一个包含130,107个词的词典。接下来，我们计算每个词的TF-IDF值，并保留TF-IDF值最高的前7,302个词。为了直观地了解这些特征（或词语），我们在表2中列出了排名前10的代表性特征。错误标记样本由 [30]人工生成，噪声率范围为0.1至0.5。我们沿用他们的做法，将错误标记率设为20%。详细结果见图2。我们比较了五种方法的性能：1）NB（错误标签），在包含错误标签的

表 2. 20 Newsgroups 数据集的 10 个代表性特征。

文档 ID	Windows	NASA	god	驱动器	苹果	IBM	car	弗吉尼亚州	MIT	空间
5554	1	1	1	1	1	1	0	1	1	1
1117	1	1	1	1	1	1	0	1	1	1
7404	1	1	0	1	1	1	0	1	1	1
2086	1	1	0	0	0	1	0	0	1	0
4770	1	1	0	0	0	1	0	0	1	1

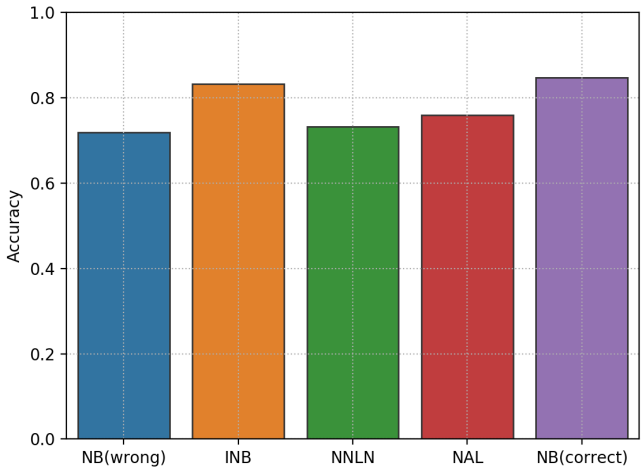

图 2. 20新闻组数据集上的分类准确率结果。

训练集上训练的NB模型；2）INB方法；3）NLNN方法；4）NAL方法；5）NB（正确标签），在标签正确的训练数据上训练的NB模型。20新闻组数据集上错误标记率为20%时的错误标记效应 Δ准确率为-12.85%。我们发现，INB方法的性能与基于正确标签的NB非常接近，优于其他方法。

直播对话数据集[3]。为了进一步说明我们的方法，我们在此展示另一项有趣的真实数据集分析，该数据集由我们的业界合作者慷慨提供。在 COVID-19 疫情期间，许多消费者转向 TikTok 等直播平台，观看主播介绍产品。消费者可以通过发送直播评论或消息向主播询问产品相关问题。及时、恰当地回复这些评论或消息，很可能会增加产品销量。然而，当大量消费者在很短时间内发送直播评论或消息时，主播要立即回复会面临很大负担。构建自动回复机器人可以减轻这种负担。一种高效自动回复机器人的可行工作流程是：先将消息分类到特定类别，然后

然后根据类别作出响应。

本研究旨在设计一个能够自动回答消费者问题的系统。为此，我们需要根据精心设计的回复策略，将这些消息分类到不同类别中；请参见 3rd表3中的列。例如，所有贷款相关问题都归为同一类别。每当一条消息被正确归入该类别时，自动系统就应按照标准流程回复消费者，相关流程列于附录4 中的图A.3。自动回复系统非常复杂，经过了大量测试和改进。因此，系统共设有 13 个类别及其对应的自动回复策略；详情请参见表3 。

随后，使用志愿者与初步设计的自动回复机器人进行交互，并收集大量消息，其中大多数与购买某一特定品牌的汽车相关。然后，要求志愿者根据预先定义的13个类别为消息添加标签。然而，我们发现这些标签并不总是正确的。出于研究目的，所有这些标签都由领域专家进行人工仔细核查。由此得到一组经过仔细核查的标签，可将其视为金标准。由此得到一个包含 $N = 1416$ 消息的数据集。对于每条消息，都提供了观测标签 Y_i 和真实标签 Y_i^*。经过仔细检查，我们发现观测标签中约有 19.49% 是错误的。

为了对分类问题建模，我们从消息的文本信息中提取一个维度为 $d = 22$ 的特征向量。每个特征都是由领域专家精心定义的二元特征，并考虑了预先定义的基于 13 个类别的自动回复策略；参见表3的最后一列。这是一个包含多个重要步骤的过程。首先，直播平台的几位领域专家为不同类别提供了基础关键词词典。接下来，使用这些基础关键词词典进行初步分类。最后，将错误预测中遗漏但重要的关键词添加到关键词词典中。由此得到最终使用的关键词集合

对于特征定义，例如，许多消费者一 re

[3]https://github.com/Helenology/Improved-Naive-Bayes-with-Mislabeld-Data/blob/main/data/5.csv

表3：消息所属的十三个类别及其对应的回复。

类别编号	类别描述	回复策略
1	贷款相关问题	询问消费者的联系信息。
2	折扣相关问题	直接回复“折扣为 XX%”。
3	汽车价格相关问题	直接回复“汽车价格为 XX 人民币”。
4	总费用相关问题	直接回复“总费用为 XX 人民币”。
5	可用性相关问题	直接回复“汽车可用/不可用”。
6	与车牌相关的问题	如果是同一省份，回答“是”；否则回答“否”。
7	与店铺地址相关的问题	直接回复店铺地址。
8	完全无关的问题	忽略这条消息，不要回复。
9	留下联系信息	直接回复“已收到消息”。
10	询问联系信息	将自动分配一名男销售员/女销售员。
11	不含汽车信息的问候消息	询问消费者的汽车偏好。
12	没有配置信息的消息	询问消费者的配置偏好。
13	关于新车或二手车的不明确消息	直接询问“您是指新车还是二手车”。

对折扣信息感兴趣。因此，相应的消息应包含“折扣”或“更便宜”等关键词。接着，可以定义一个二元特征 X_{17}，用以表示消费者是否对折扣信息感兴趣。可以看出，这一特征不仅对后续的分类任务有用，而且也易于解释。为便于直观理解，直播对话数据集中所有这些特征的详细描述列于表 4。共有 22 个相互独立的二元变量。如果某条消息满足表 2nd 中 4列所列出的条件，则相应变量取值为 1，否则为 0。

实验设置。我们将80%的数据随机分配到训练集，其余20%分配到测试集。训练集包含正确和错误标签。然而，测试集仅包含正确标注的数据，用于评估分类准确率。训练集和测试集的划分重复进行 $B = 100$ 次。在每次划分下，我们比较五种方法的分类准确率：1）NB（错误标签），即使用包含错误标签的训练集训练的NB模型；2）INB方法；3）NLNN方法；4）NAL方法；5）NB（正确标签），即使用正确标签的训练集训练的NB模型。

实验结果。错误标记效应对直播对话数据集 的影响为 -7.32%。图 3 展示了五种方法在 $B = 100$ 数据集划分上获得的分类准确率结果。所提出的 INB 方法优于其他方法。平均而言，通过加入错误标记机制，INB 方法使 NB 方法的分类准确率提高了 3.7%。INB 方法的预测性能与 NB（正确标签）方法非常接近，后者是在标签全部正确的数据上训练的 NB 方法。

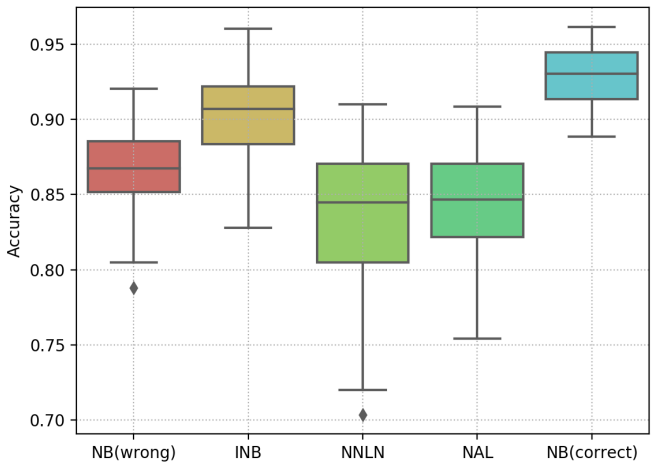

图3. 直播对话数据集上的分类准确率结果。

5. 讨论

我们提出了一种用于文本分类的 INB 方法。该方法在分析上简单明了，无需对正确标签和错误标签进行主观判断。通过指定错误标签的生成机制，我们使用期望最大化算法优化相应的对数似然函数。我们通过模拟实验说明了所提出的 INB 方法的优势。此外，我们还提供了一个直播对话数据集，用于研究 INB 方法的实证性能。数值结果均表明，INB 方法在存在错误标注数据的数据集上具有竞争力。

最后，我们在此讨论一个值得未来研究的有趣课题。在这项工作中，我们专注于具有二元特征的经典朴素贝叶斯方法。另一方面，连续变量在实际应用中也很常见。因此，

表 4. 自变量说明。

变量名称	实际含义
X_1	该消息是否仅包含汽车信息?
X_2	该消息是否为疑问句?
X_3	该消息是否为消费者发送的第一条消息?
X_4	这条消息是否是关于某一辆特定汽车的?
X_5	消息中是否提供了详细的汽车配置信息?
X_6	消息中是否包含配置信息?
X_7	这条消息是否是关于汽车店地址的?
X_8	消息中是否提供了消费者的联系信息?
X_9	这条消息是否是关于新车的?
X_{10}	这条消息是否与二手汽车有关?
X_{11}	消费者是否在这条消息中询问联系信息?
X_{12}	这条消息是否是在陈述某一辆特定汽车的情况?
X_{13}	消费者是否在之前的消息中留下了联系信息?
X_{14}	这条消息是否是在询问车牌?
X_{15}	这条消息是否是在询问总费用?
X_{16}	这条消息是在询问汽车价格吗?
X_{17}	这条消息是在询问折扣吗?
X_{18}	这条消息是在询问汽车是否有货或是否需要预订吗?
X_{19}	这条消息是在询问贷款吗?
X_{20}	这条消息不是在询问贷款、总费用、汽车价格、折扣或索取 联系信息、留下联系信息或可用性吗? 汽车店地址还是车牌?
X_{21}	这条消息是否与贷款、总费用、汽车价格、折扣或询问 联系信息、留下联系信息、是否有货, 汽车店地址还是车牌? 这条消息是消费者发送的第一条消息吗?
X_{22}	这条消息是否与贷款、总费用、汽车价格、折扣或询问 联系信息、留下联系信息、可用性, 汽车店地址或车牌? 这条消息是消费者发送的第一条消息吗? 我们能否判断消费者在这条消息中指的是哪辆汽车?

如何将连续特征纳入所提出的 INB 框架，便成为一个备受关注的问题。这里的关键在于，应对连续特征施加何种分布假设。一种直接的解决方案是采用高斯假设。在这种情况下，可以很容易地开发出一种 EM 类算法（详情请见附录A.1 ）。然而，这是否是实现最佳实证性能的最优假设，目前尚不明确。更灵活的非参数分布假设或许是有前景的替代方案。对此显然还需要进行更深入的分析。

附录

A.1 连续特征

这里我们考虑如何在高斯假设下向 INB 模型添加连续特征。假设有d_1 个二元特征和 d_2 个连续特征。用 $X_i = (X_{i1}, \cdots, X_{id_1})^\top$表示二元特征。用 $Z_i = (Z_{i1}, \cdots, Z_{id_2})^\top$表示连续特征。我们将所有$X_i$收集到 $\mathbb{X}$ 中，并将所有 Z_i收集到 $\mathbb{Z}$中。与高斯朴素贝叶斯

在贝叶斯模型中，我们假设 $p(z_{ij}|Y_i^* = k) = \phi_{jk}(z_{ij})$对于 $1 \le k \le K$ 和 $1 \le j \le d_2$，其中 $\phi_{jk}(\cdot)$是均值为 μ_{jk} 、标准差为 σ_{jk}的正态分布密度函数。整个参数集变为 $\theta = \{p_{jk}, 1 \le j \le d_1, 1 \le k \le K\} \cup \{\pi_k, 1 \le k \le K\} \cup \{\rho_{k_1 k_2}, 1 \le k_1, k_2 \le K\} \cup \{\mu_{jk}, 1 \le j \le d_2, 1 \le k \le K\} \cup \{\sigma_{jk}, 1 \le j \le d_2, 1 \le k \le K\}$。然后，相应的对数似然变为

$$
\begin{aligned}
\ell_c(\theta) &= \ln P(\mathbb{X}, \mathbb{Z}, \mathbb{Y}, \mathbb{Y}^*|\theta) \\
&= \sum_{i=1}^{N} \ln P(Y_i^*|\theta) + \sum_{i=1}^{N} \ln P(Y_i|Y_i^*, \theta) \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{d_1} \ln P(X_{ij}|Y_i^*, \theta) + \sum_{i=1}^{N}\sum_{j=1}^{d_2} \ln P(Z_{ij}|Y_i^*, \theta) \\
&= \sum_{i=1}^{N} \ln \pi_{Y_i^*} + \sum_{i=1}^{N} \ln \rho_{Y_i Y_i^*} + \sum_{i=1}^{N}\sum_{j=1}^{d_1} X_{ij} \ln p_{jY_i^*} \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{d_1} (1 - X_{ij}) \ln (1 - p_{jY_i^*}) + \sum_{i=1}^{N}\sum_{j=1}^{d_2} \ln \phi_{jY_i^*}(Z_{ij}).
\end{aligned}
$$

表 5. 不同 ρ_{kk} 间隔和不平衡样本量下不同方法的有限样本性能。

ρ_{kk} 区间	n	均方误差×10^{-3} NB	均方误差×10^{-3} INB	准确率 (%) NB	准确率 (%) INB	准确率 (%) NB-T	准确率 (%) NLNN	准确率 (%) NAL	曲线下面积 (%) NB	曲线下面积 (%) INB	曲线下面积 (%) NB-T	曲线下面积 (%) NLNN	曲线下面积 (%) NAL	Δ准确率 (%)
[0.55, 0.65)	1000	2.5	1.6	72.4	90.7	97.0	23.5	47.0	92.3	99.1	99.9	56.3	68.0	-19.4
	5000	1.3	0.3	88.9	95.2	96.2	38.1	78.8	98.5	99.6	99.8	64.2	94.4	-6.4
	10000	1.2	0.2	92.0	96.4	96.5	52.3	89.2	99.4	99.8	99.8	76.3	98.3	-4.3
[0.65, 0.75)	1000	2.1	1.6	81.5	90.8	97.0	27.6	50.1	96.5	99.1	99.9	58.4	73.9	-10.2
	5000	0.9	0.3	92.3	95.2	96.2	44.0	84.3	99.3	99.6	99.8	66.8	96.7	-3.0
	10000	0.8	0.2	94.1	96.4	96.5	61.4	91.6	99.6	99.8	99.8	77.6	99.0	-2.3
[0.75, 0.85)	1000	1.8	1.6	86.5	90.9	97.0	32.1	52.4	97.9	99.1	99.9	60.5	77.4	-5.3
	5000	0.6	0.3	94.0	95.2	96.2	51.1	87.7	99.5	99.6	99.8	68.7	97.9	-1.3
	10000	0.5	0.2	95.3	96.3	96.5	61.5	93.0	99.7	99.8	99.8	74.1	99.3	-1.1
[0.85, 0.95)	1000	1.5	1.5	90.2	91.3	97.0	33.7	56.3	98.8	99.2	99.9	60.7	80.1	-1.5
	5000	0.4	0.3	94.8	95.2	96.2	47.6	89.4	99.6	99.6	99.8	66.4	98.4	-0.4
	10000	0.2	0.2	96.1	96.3	96.5	62.9	94.1	99.8	99.8	99.8	75.5	99.5	-0.3
[1.0, 1.0]	1000	1.4	1.5	91.8	91.7	97.0	37.4	58.5	99.3	99.3	99.9	61.5	81.3	0.0
	5000	0.3	0.3	95.3	95.3	96.2	47.4	91.4	99.6	99.6	99.8	64.6	98.9	0.0
	10000	0.1	0.1	96.4	96.4	96.5	59.4	95.1	99.8	99.8	99.8	72.1	99.7	0.0

相应地，我们可以开发一种与小节3.5中类似的期望最大化算法。

E步。在第 t次迭代中，给定观测数据（$\mathbb{X}, \mathbb{Z}, \mathbb{Y}$）和当前参数估计 $\widehat{\theta}^{(t)}$，我们计算对数似然 $\ell_c(\theta)$的期望。条件期望可推导为

$$Q\left(\theta, \widehat{\theta}^{(t)}\right) = E\left\{\ell_c(\theta)\middle|\mathbb{X}, \mathbb{Z}, \mathbb{Y}, \widehat{\theta}^{(t)}\right\}$$
$$= \sum_{i=1}^{N}\sum_{k=1}^{K}\widehat{\gamma}_{ik}^{(t)}\ln\pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K}\widehat{\gamma}_{ik}^{(t)}\ln\rho_{Y_i k} + \sum_{i=1}^{N}\sum_{j=1}^{d_1}\sum_{k=1}^{K}X_{ij}\widehat{\gamma}_{ik}^{(t)}\ln p_{jk}$$
$$+ \sum_{i=1}^{N}\sum_{j=1}^{d_1}\sum_{k=1}^{K}(1-X_{ij})\widehat{\gamma}_{ik}^{(t)}\ln(1-p_{jk}) + \sum_{i=1}^{N}\sum_{j=1}^{d_2}\sum_{k=1}^{K}\widehat{\gamma}_{ik}^{(t)}\ln\phi_{jk}(Z_{ij}),$$

其中，$\widehat{\gamma}_{ik}^{(t)}$ 表示 $1 \le i \le N$和$1 \le k \le K$对应的 $P(Y_i^* = k|X_i$、Z_i、Y_i、$\widehat{\theta}^{(t)})$。为计算 $\widehat{\gamma}_{ik}^{(t)}$，我们首先定义 $\widehat{\zeta}_{ik}^{(t)}$ 如下。

$$\widehat{\zeta}_{ik}^{(t)} = \widehat{\pi}_k^{(t)}\widehat{\rho}_{Y_i k}^{(t)}\left[\prod_{j=1}^{d_1}\widehat{p}_{jk}^{(t)X_{ij}}\left\{1-\widehat{p}_{jk}^{(t)}\right\}^{1-X_{ij}}\right] \times \left[\prod_{j=1}^{d_2}\widehat{\phi}_{jk}^{(t)}(Z_{ij})\right],$$

其中 $\widehat{\phi}_{jk}^{(t)}(Z_{jk}) = 1/\sqrt{2\pi\widehat{\sigma}_{jk}^{(t)2}} \times \exp\{-(Z_{jk} - \widehat{\mu}_{jk}^{(t)})/(2\widehat{\sigma}_{jk}^{(t)2})\}$。然后我们计算 $\widehat{\gamma}_{ik}^{(t)} = \widehat{\zeta}_{ik}^{(t)} / \sum_{k=1}^{K}\widehat{\zeta}_{ik}^{(t)}$。

M步。在此步骤中，我们最大化条件期望 $Q(\theta, \widehat{\theta}^{(t)})$，以获得新的估计值 $\widehat{\theta}^{(t+1)}$。注意，对于 $1 \le k_2 \le K$，$\sum_{k=1}^{K}\pi_k = 1$ 和 $\sum_{k_1=1}\rho_{k_1 k_2} = 1$。基于拉格朗日乘数法，我们可以将拉格朗日函数定义为

$$Q = Q\left(\theta, \widehat{\theta}^{(t)}\right) + \lambda\left(1 - \sum_{k=1}^{K}\pi_k\right) + \sum_{k_2=1}^{K}\lambda_{k_2}\left(1 - \sum_{k_1=1}\rho_{k_1 k_2}\right),$$

其中，λ 和 λ_{k_2} 对于 $1 \le k_2 \le K$ 而言，是拉格朗日乘数法中使用的额外参数。在最大化拉格朗日函数 Q后，第（$t+1$）次迭代中的更新参数计算如下：

$$\widehat{\pi}_k^{(t+1)} = \sum_{i=1}^{N}\widehat{\gamma}_{ik}^{(t)}/N,\ 1 \le k \le K,$$
$$\widehat{p}_{jk}^{(t+1)} = \left(\sum_{i=1}^{N}X_{ij}\widehat{\gamma}_{ik}^{(t)}\right) \Big/ \left(\sum_{i=1}^{N}\widehat{\gamma}_{ik}^{(t)}\right),\ 1 \le j \le d_1,\ 1 \le k \le K,$$
$$\widehat{\rho}_{k_1 k_2}^{(t+1)} = \left(\sum_{i=1}^{N}I(Y_i = k_1)\widehat{\gamma}_{ik_2}^{(t)}\right) \Big/ \left(\sum_{i=1}^{N}\widehat{\gamma}_{ik_2}^{(t)}\right),\ 1 \le k_1, k_2 \le K,$$
$$\widehat{\mu}_{jk}^{(t+1)} = \left(\sum_{i=1}^{N}Z_{ij}\widehat{\gamma}_{ik}^{(t)}\right) \Big/ \left(\sum_{i=1}^{N}\widehat{\gamma}_{ik}^{(t)}\right),\ 1 \le j \le d_2,\ 1 \le k \le K,$$
$$\left(\widehat{\sigma}_{jk}^{(t+1)}\right)^2 = \left\{\sum_{i=1}^{N}\widehat{\gamma}_{ik}^{(t)}\left(Z_{ij} - \widehat{\mu}_{jk}^{(t+1)}\right)^2\right\} \Big/ \left(\sum_{i=1}^{N}\widehat{\gamma}_{ik}^{(t)}\right).$$

适用于二元和连续特征的期望最大化算法，与小节 3.5 中的算法类似。由于我们有连续特征 Z_is，更新 $\widehat{\gamma}_{ik}^{(t)}$ 涉及参数（$\widehat{\mu}_{jk}^{(t)}$，$\widehat{\sigma}_{jk}^{(t)}$），对应于 $1 \le j \le d_2$和 $1 \le k \le K$。此外，在 M 步中，我们还需要更新（$\widehat{\mu}_{jk}^{(t)}$，$\widehat{\sigma}_{jk}^{(t)}$）。

A.2 不平衡样本量

接下来，我们研究不平衡样本量的影响。为此，我们沿用第4.1 节相同的数据生成策略，但只做两处修改。第一处修改是将类别 1 的先验概率设为其他类别概率的 3 倍。也就是说，$\pi_1 = 0.428$且 $\pi_k = 0.143$ 对于 $k > 1$。第二处修改是将数据规模设为 $n = 1000$、5000 和 10000。我们将类别设为 $K = 5$。为进行验证，每次随机重复中都会保留约 20% 的模拟数据用于测试。详细结果见表

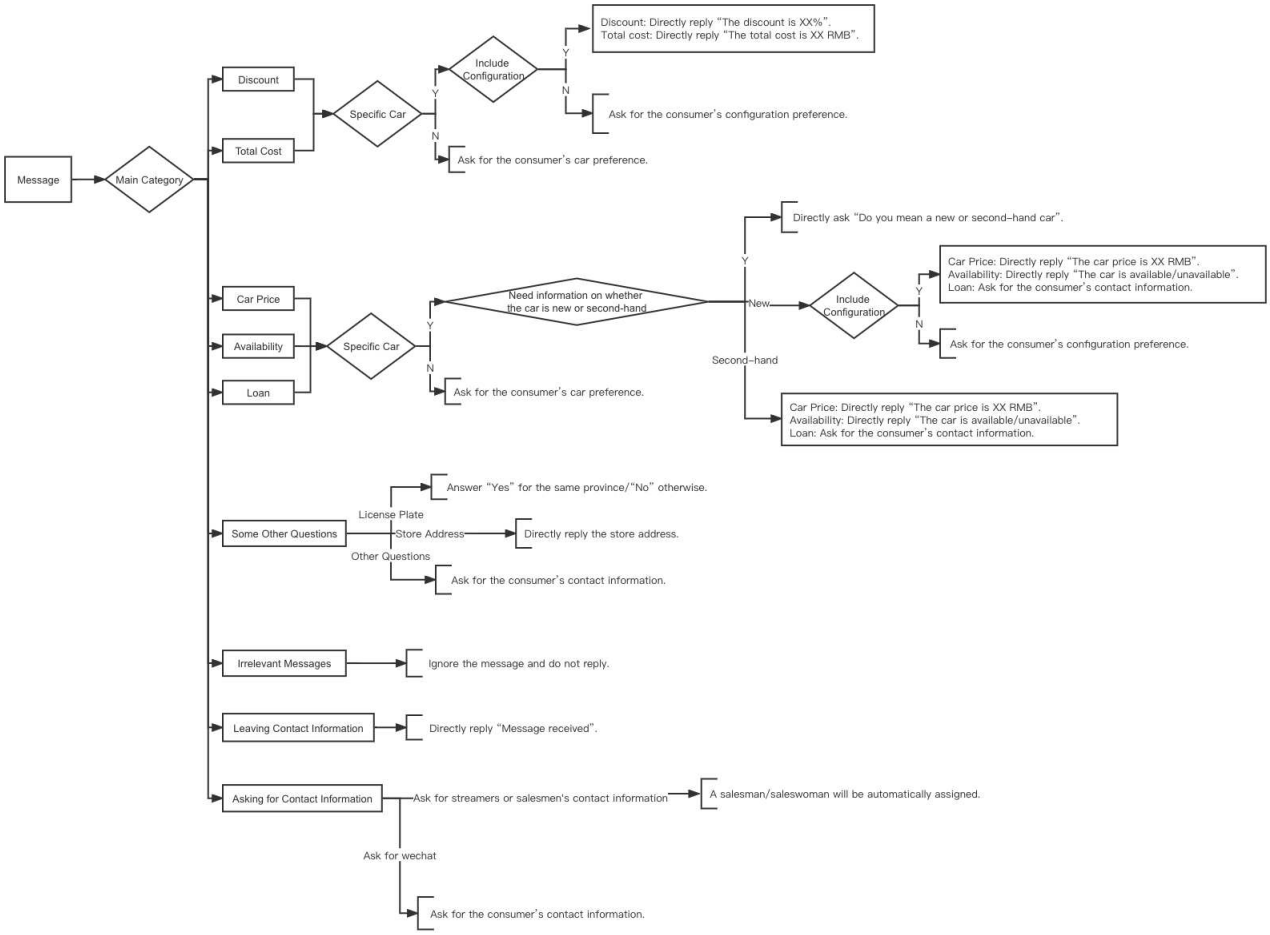

图4. 回复流程图。

5。我们发现，结果在定性上与表1中的结果相似。

A.3 标注过程流程图

为了自动回复消费者的问题，我们首先根据精心设计的回复策略，将这些消息归入不同类别；参见 3rd 表3中的列。每当一条消息被正确归入某个类别后，便会根据图4所列流程回复消费者。例如，如果消费者询问某款汽车的折扣，但未提供配置信息，那么他或她会收到一条询问其配置偏好的消息。

致谢

朱雪宁的研究得到国家自然科学基金委员会（NSFC, 11901105、71991472、U1811461）、上海市青年科技英才扬帆计划（19YF1402700）以及复旦大学数据科学学院复旦-新再灵大数据联合研究中心的资助。王菲菲王菲菲的研究得到国家自然科学基金委员会（72001205）和中国国家统计科学研究项目（2022LD06）的资助。王汉生的研究部分得到国家自然科学基金委员会（编号：11831008）的资助，另有部分得到统计与数据科学高级理论与应用重点实验室开放研究基金（KLATASDS-MOE-ECNU-KLATASDS2101）的资助。

2022年9月3日收稿

参考文献

[1] Ahmad, I. (2019)。含噪标注训练数据上的分类器性能：手写数字分类任务的实证研究。载于 *International Work-Conference on Artificial Neural Networks* 414 425 – 。Springer。[2] Bekker, A. J. 和 Goldberger, J. (2016)。基于不可靠标签训练深度神经网络。载于 *2016 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* 2682 2686 - , - 。

[3] Brodley, C. E.和 Friedl, M. A. (1999)。识别错误标注的训练数据。*Journal of Artificial Intelligence Research* **11** 131– 167。

[4] Buhrmester, M., Kwang, T. 和 Gosling, S. D. (2011年)。亚马逊 Mechanical Turk：一种新的低成本、高质量数据来源? *Perspectives on Psychological Science* **6** 3-5。PubMed唯一标识符: 26162106。[5] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu Krishnan Pan Baldan Beijbom , Q., , A., , Y., , G. 和 , O. (2020年)。nuScenes：一个用于自动驾驶的多模态数据集。载于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 。

[6] Cognilytica, R. (2019年)。*Data Engineering, Preparation, and Labeling for AI 2019: Getting Data Ready for Use in AI and Machine Learning Projects* 。[7] Dirican, C. (2015年)。机器人学、人工智能对商业和经济学的影响。*Procedia - Social and Behavioral Sciences* **195** 564 573 - 。世界技术、创新与创业大会。[8] Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A. 和 Strachan, R. (2014年)。用于多类分类任务的混合决策树和朴素贝叶斯分类器。*Expert Systems With Applications* **41** 1937 1946 – 。[9] Fawcett, T. (2006 *Pattern*)。ROC分析简介。*Recognition Letters* **27** 861 874 – 。[10] Goldberger, J. 和 Ben-Reuven, E. (2017年)。使用噪声适应层训练深度神经网络。载于 *International Conference on Learning Representations* 。[11] Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. 和 Sugiyama, M. (2018年)。协同教学：使用极度含噪标签对深度神经网络进行鲁棒训练。载于 *Advances in Neural Information Processing Systems* S. Bengio (, , , 和 lach H. Larochelle K. Grauman N. Cesa-Bianchi R. Garnett, 编者) 31。Curran Associates, Inc.[12] Jeatrakul, P., Wong, K. W. 和 Fung, C. C. (2010年)。使用误分类分析进行分类数据清洗。*Journal of Advanced Computational Intelligence and Intelligent Informatics* **14** 297 302 – 。[13] Ji, D., Oh, D., Hyun, Y., Kwon, O.-M. 和 Park, M.-J.(2021年)。如何从不确定性中处理含噪标签以实现鲁棒学习。*Neural Networks* **143** 209-217。[14] Jiang, L., Zhou, Z., Leung, T., Li, L.-J. 和 Fei-Fei, L. (2018年)。MentorNet：在损坏标签上为极深神经网络学习数据驱动课程。载于 *Proceedings of the 35th International Conference on Machine Learning* J. Dy (和 A. Krause, 编者)。**80** 2304 2313 *Proceedings of Machine Learning Research* – 。PMLR。[15] Joskowicz, L., Cohen, D., Caplan, N. 和 Sosna, J.(2019年)。CT中结构手工轮廓勾画的观察者间变异性。*European Radiology* **29** 1391–1399。[16] Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D. 和 Brossard, E. (2016年)。MegaFace基准：用于大规模识别的100万张人脸。载于 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 4873 4882 - 。[17] Lang, K. (1995年)。NewsWeeder：学习过滤网络新闻。载于 *Proceedings of the Twelfth International Conference on Machine Learning* 331 339 - 。[18] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan Dollár Zitnick 2014 , D., , P. 和 , C. L. ( )。Microsoft COCO：情境中的常见物体。载于 *European Conference on Computer Vision* 740 755 – 。Springer。[19] Malach, E. 和 Shalev-Shwartz, S. (2017年)。将"何时更新"与"如何更新"解耦。载于 *Advances in Neural Information Processing Systems* I. Guyon U. V. Luxburg (, , S. Ben- , , , 和 gio H. Wallach R. Fergus S. Vishwanathan R. Gar- , 编者) 。 nett **30** Curran Associates, Inc.[20] Masi, I., Yue, W., Hassner, T. 和 Natarajan, P. (2018年)。深度人脸识别：综述。载于 *2018 31st SIBGRAPI Conference on Graphics Patterns and Images (SIBGRAPI)* , 。

[21] Northcutt, C., Athalye, A. 和 Mueller, J. (2021年)。测试集中的普遍标签错误会破坏机器学习基准的稳定性。载于 *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (J. Vanschoren 和 S. Yeung, 编者) 1.[22] Patrini, G., Rozza, A., Menon, A. K., Nock, R. 和 Qu, L. (2017年)。使深度神经网络能够抵御标签噪声：一种损失校正方法。载于 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2233 2241 - 。[23] Popenici, S. A. D. 和 Kerr, S. (2017年)。探索人工智能对高等教育教学与学习的影响。*Research and Practice in Technology Enhanced Learning* **12** 。

[24] Ramakrishnan, G., Chitrapura, K. P., Krishnapuram, R. 和 Bhattacharyya, P. (2005年)。一种处理文本分类中近似、不准确或不完整标注的模型。载于 *Proceedings of the 22nd International Conference on Machine Learning* 681 688 – 。

[25] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma Huang Karpathy Khosla Bernstein 、 S., , Z., , A., , A., , M., Berg Fei-Fei 2015 、 A. C. 和 , L. ( )。ImageNet大规模视觉识别挑战赛。*International Journal of Computer Vision (IJCV)* **115** 211 252 - 。[26] Smyth, P., Fayyad, U., Burl, M., Perona, P. 和 Baldi, P. (1994年)。根据对金星图像的主观标注推断真实标注。载于 *Advances in Neural Information Processing Systems* G. Tesauro D. Touretzky T. Leen **7** (, 和 , 编者)。麻省理工学院出版社。[27] Thongkam, J., Xu, G., Zhang, Y. 和 Huang, F. (2008年)。用于乳腺癌生存预测中异常值检测的支持向量机。载于 *Advanced Web and Network Technologies and Applications* Y. Ishikawa J. He G. Xu Y. Shi , (, , , , G. Huang C. Pang Q. Zhang G. Wang 99 109 , , 和 , 编者) – 。Springer Berlin Heidelberg, 德国柏林。[28] Ting, S., Ip, W., Tsang, A. H. 等 (2011年)。朴素贝叶斯是文档分类的良好分类器吗? *International Journal of Software Engineering and Its Applications* **5** 37–46。[29] Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu Liu Sun Yang Liu Chen Zhou 、 Y., , X., , H., , R., , X., , J., , H., Ben Ayed Zheng 2021 、 I. 和 , H. ( )。用于医学图像自动分割的标注高效深度学习。*Nature Communications* **12** 5915 。[30] Wang, W., Guo, B., Shen, Y., Yang, H., Chen, Y. 和 Suo, X. (2021 *Machine Learning* **110** 907 931)。标签噪声下的鲁棒监督主题模型。– 。[31] Wu, X.-Z. 和 Zhou, Z.-H. (2017年)。多标签性能度量的统一视角。载于 *Proceedings of the 34th International Conference on Machine Learning* D. Precup Y. W. Teh (和 , 编者)。*Proceedings of Machine Learning Research* **70** 3780–3788。PMLR。[32] Zhang, C., Bengio, S., Hardt, M., Recht, B. 和 Vinyals, O. (2017年)。理解深度学习需要重新思考泛化。载于 *International Conference on Learning Representations* 。[33] Zhang, Y., Niu, G. 和 Sugiyama, M. (2021年)。通过全变差正则化，仅从含噪标签中学习噪声转移矩阵。载于 *Proceedings of the 38th International Conference on Machine Learning* M. Meila T. Zhang (和 , 编者)。*Proceedings of Machine Learning Research* **139** 12501–12512。PMLR。[34] Zhu, H. (2020年)。用于药物发现的大数据与人工智能建模。*Annual Review of Pharmacology and Toxicology* **60** 573 589 – 。

Qianhan Zeng 光华管理学院，北京大学 电子邮箱：helenology@stu.pku.edu.cn

Yingqiu Zhu 统计学院、对外经济贸易大学 电子邮箱：rozen0maiden@126.com

朱雪宁 数据科学学院，复旦大学 电子邮箱：xueningzhu@fudan.edu.cn

王菲菲，中国人民大学统计学院，电子邮箱：feifei.wang@ruc.edu.cn

Weichen Zhao，河海大学计算机与信息学院，电子邮箱：zhaoweichen@hhu.edu.cn

Shuning Sun 数据科学学院，复旦大学 电子邮箱：snsun21@m.fudan.edu.cn

苏萌 北京百分点科技集团有限公司 电子邮箱：meng.su@percent.cn

王汉生 北京大学光华管理学院 电子邮箱：hansheng@pku.edu.cn