

研究提案创新案例——分类器校准

爱德华·米尔森

1. 执行摘要

机器学习 (ML) 已经发现了广泛的应用,包括在安全关键领域,例如医疗诊断1、2和自动驾驶汽车/驾驶辅助系统 3、4、5,并且预计只会进一步增长,特别是在医疗保健领域必须有更高的准确性和系统安全性。如果这些领域的伤害或死亡率增加,则必须确保这些系统的准确性。

分类器是一种算法,它接收一个数据点并为其分配一个标签。一个重要的例子是识别图像中的对象。许多分类器还给出了一个不确定性,这应该是预测标签正确的概率。然而,许多分类器给出了未校准的不确定性7、8,这本质上意味着不确定性并不代表真实的概率。因此,这些不确定性意义不大,难以解释,因此在用于决策过程时不能被视为可靠。我们可以使用校准器在许多场景中获得更准确的概率,但它们在现实世界数据集上的表现往往不能令人满意。

我们提出了一个项目,该项目将分析当前最先进的校准器,评估它们在许多现实世界数据集上的性能,并为这些技术提供可能的改进。更重要的是,我们将提出全新的校准方法,经过改进后可以与现有的校准器进行比较。如果该项目成功,我们将看到使用我们的新技术在分类器校准方面的显着改进,这些技术可以很容易地在生产 ML 系统中实施,从而提供即时收益。这将涉及招募一小群博士生,他们每个人都将探索不同的方向作为他们研究的一部分,以及一名博士后研究人员在最有前途的方向上进行更集中的分析。我们将与 Facebook 的 AI 研究实验室合作,将我们的新校准技术引入广泛使用的机器学习库 PyTorch,从而快速跟踪它们的应用,从而对现实世界的问题产生直接影响。

2 潜在机会评估

2018 年,Elaine Herzberg 成为世界上第一个在与自动驾驶汽车相撞中丧生的人。⁹这一事件阻碍了在不久的将来广泛使用自动驾驶汽车的努力,但它也带来了注意在高风险情况下使用“黑盒”机器学习模型的风险。有人认为,在这种关键情况下,我们必须转向使用更具可解释性的模型。¹⁰机器学习,尤其是深度学习,在过去十年中已成为许多领域不可或缺的一部分,¹¹并有望在未来继续增长。^{future.6}特别是一类算法,称为分类器,由于深度学习的兴起,得到了极大的改进。

分类器已经开始在医疗诊断系统^{1、2}和自动驾驶系统等安全关键场景中使用。^{3、4、5}如果这些系统性能欠佳,则可能造成伤害、死亡或财产损失。

一些前沿模型的一个关键问题,特别是在深度学习中,是它们没有为他们的估计提供不确定性。如果没有不确定性,就无法知道模型对其预测的信心程度,因此很难信任它们做出重要决策。一直在推动确保模型在未来更具概率性。¹²

然而,一些模型确实已经提供了不确定性。这就提出了一个关键问题:我们如何知道这些

分数（我们通常将一个类的输出概率称为分数）表示预测正确的真实概率？正确性的一个概念是要求校准输出的不确定性。

这意味着，对于任何特定分数，实际属于预测类的具有该分数的实例的真实比例与该分数相匹配。例如，如果有一组图像都被一个识别猫的分类器给出了 70% 的分数，那么这些图像中恰好有 70% 应该实际上包含一只猫。

尽管校准不是准确概率的完美度量（您可以构建分类器经过技术校准的示例，即使它只是为任何输入输出相同的分数），但在正常情况下，它是一个非常理想的属性。

许多模型不输出校准后的概率。⁷在这些情况下，我们可以使用事后校准器，它尝试学习分数和真实概率之间的映射。已经提出了一些不同的校准器，包括 Platt 标度、等渗校准器，以及最近的 Beta 和 Dirichlet 校准器。⁷都有自己的缺点，我们距离一刀切的解决方案还有很长的路要走问题。

要了解目前广泛使用的校准技术，我们可以查看一些流行的机器学习库。Scikit-learn¹³是一个 Python 库，主要关注传统的机器学习技术（即不是深度神经网络）。我们从文档中看到 Scikit-learn 支持逻辑校准（Platt scaling）和等渗校准。这些都是二元分类器（只能识别一个类别的分类器，例如“猫”与“非猫”）的校准技术。二进制校准器不能很好地推广到多类案例⁷（可以识别多个类别的分类器，例如“Cat”、“Dog”、“Horse”）。此外，这两种技术现在已经很老了，特别是 Platt 标度最近已经推广到使用称为 Beta 校准的技术来限制较少的假设。¹⁵

PyTorch¹⁶是一个流行的深度学习库。大多数时候，深度神经网络被用作多类分类器。尽管神经网络输出的分数是非标准化的，但通常使用一种称为 softmax 的技术来产生伪概率。然而，这些伪概率通常是未经校准的，¹⁷并且 PyTorch 没有为多类校准提供本机实现。专门用于神经网络的技术，特别是温度缩放¹⁷，自从它们被提出以来已经越来越流行，尽管在 PyTorch 的情况下，用户仍然必须手动实现它们。

如前所述，ML 预计只会在未来变得更加融入我们的生活，因此虽然校准可能并不迷人，但至关重要的是，我们使用的模型能够在做出关键决策时为我们提供准确的概率。对社会的潜在好处是巨大的，但现在需要开始工作以确保广泛采用。

3 价值主张

Dirichlet 校准是在最近的一篇论文中提出的。⁷它是 Beta 校准的多类推广，也可以看作是温度标定的限制较少的版本。在论文中，作者证明了 Dirichlet 校准使用各种评估指标产生了最先进的性能。然而，他们自己承认 Dirichlet 校准只是一个起点，在生产基于指数族的不同概率分布的新校准器方面还有更多的研究工作要做（Dirichlet 校准本身基于 Dirichlet 分布，指数族的成员），甚至基于分布的混合。

图 1 显示了两个数据集的可靠性图：一个 Dirichlet 校准表现良好，一个表现不佳。可靠性图简单地显示了预测分数与真实阳性比例的关系。从角到角向上的对角直线是理想的，因为那时我们的预测分数与真实概率相匹配。我们可以在图 1a 中看到，在这个数据集上，校准

曲线变成一条直线,这意味着这个分类器现在已经校准,因此它的不确定性是可靠的。然而在图 1b 中,我们看到即使在对其他数据集应用 Dirichlet 校准之后,分类器的校准仍然很差。这意味着我们不能依赖它的不确定性,因此我们不应该在安全关键的情况下使用它。显然,Dirichlet 校准并非在所有情况下都是完美的解决方案。

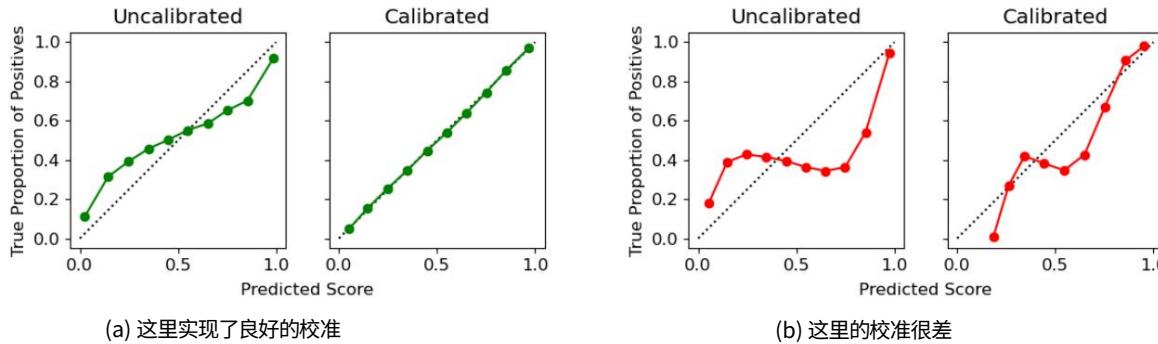


图 1:显示 Dirichlet 校准在某些数据集上表现不同的可靠性图。直线对角线是最佳校准曲线。

这就是我们提出的研究项目的用武之地。我们的目标是研究新的方向,例如我们将与之合作的狄利克雷校准论文的作者所建议的方向。具体来说,我们将识别 Dirichlet 校准性能不佳的数据集(不可能提供“好”或“差”性能的精确定义,但有一系列评估指标可用于比较校准器,例如 Brier 分数、置信度预期校准误差(ECE)、分类 ECE、最大校准误差等)。一旦确定了这一点,我们就可以开始研究替代校准图。受 Dirichlet 校准的启发,我们已经计划研究从指数族概率分布(如 Gamma 和 Wishart 分布)以及混合分布派生的校准图。

我们还将研究最近提出的称为 Metalog 分布的概率分布。¹⁸如果认为它在校准中的使用被认为是可行的,那么 Metalog 分布由于其高度的灵活性可能非常有前途,允许它对更广泛的分数分布进行建模,这将产生更好的校准图,因此得到更好的校准概率。此外,Metalog 分布可以使用线性最小二乘法进行拟合,与用于许多其他分布的基于下降的方法相比,这是一种非常有效的算法,这意味着需要更少的计算资源。这具有通过使用更少的能源来减少我们对环境的影响的额外好处。

在与开发 PyTorch 的 Facebook 人工智能研究实验室(FAIR)的讨论中,他们表示希望通过包含校准技术来扩展其库的功能,尽管他们认为当前的选择是有限的。作为对此的回应,他们愿意为我们的研究项目提供部分资金,以便他们有更好的技术来实施。他们认为他们没有必要获得知识产权,只要由此产生的作品属于公共领域,以防止任何竞争对手开发类似的专有系统。由于校准是一个相当小众的研究领域,他们为我们已经涉及多名专家的项目提供资金更容易和更便宜,而不是通过昂贵的过程聘请新的研究人员和组装他们自己的项目。他们还希望这种伙伴关系能够让他们至少招募一名博士生,到项目结束时,他们也将成为专家。

PyTorch 在深度学习中被广泛使用,因此原生校准实现将使所有在该领域工作的人受益,并将鼓励更多人在开发实际使用的系统时考虑校准的概念。

4 影响计划

该项目将与 Dirichlet 校准论文的作者合作进行。⁷ 我们已经在与他们进行讨论，我们都渴望使该项目成为现实。此外，计划招收3-4名新博士生开展研究主体，1名博士后研究员。

Facebook 已经为该项目承诺了 500,000 英镑，因此我们将需要 500,000 英镑的赠款来支付员工的薪水以及博士生的津贴。除了大学必须支付的标准维护成本外，赠款还将支付设备成本，由于项目的性质，设备成本将很低，主要限于研究人员的个人电脑。项目工期为4年。

到项目结束时，我们将完成 3 个阶段/目标。首先，将深入分析现有校准技术的优势和劣势，我们的研究结果通过科学论文发表在知名会议上，例如 NeurIPS 和 ICML。其次，将在实验环境中衍生、实施和测试各种新颖的校准技术。这些方法中最有希望的将进入第三阶段，在那里它们将被开发成一个名为“calibPy”的公开可用的开源库，作为技术的演示。之后，Facebook 打算将这些技术整合到 PyTorch 中。这三个阶段的详细信息将在以下段落中列出。

在第一阶段，被称为“评估”阶段，研究人员将考虑许多现有的校准器，包括用于二元分类器的 Platt 缩放、等渗校准、Beta 校准和经验分箱，以及用于二进制分类器的温度缩放、矢量缩放、矩阵缩放和 Dirichlet 校准。多类分类器。

他们将在许多最常见的数据集上进行评估，例如 MNIST、CIFAR-10、Wine Quality 和 Pima Indians Diabetes，使用范围广泛的评估指标和适当的评分规则，包括 Binary ECE、Binary MCE、Confidence ECE、Classwise ECE、Brier 分数和对数损失。研究结果将通过在知名会议上发表的科学论文进行传播。

在被称为“探索”阶段的第二阶段，我们将探索上一节中讨论的途径，即使用不同的指数族概率分布、Metalog 分布，以及研究如何将混合分布应用于校准问题。这些调查将以非正式的方式进行，研究小组将定期开会，分享调查结果并确定哪些方向最具潜力。

在被称为“细化”阶段的最后阶段，我们将采用与第一阶段类似的方法，从探索阶段获取最佳想法并更彻底地测试它们。那些被认为是最先进的校准方法将被开发成一个开源库“calibPy”，并将撰写科学论文详细介绍它们的功能和性能。这些论文和图书馆都将使其他研究人员能够评估我们的新方法，希望他们能够成功现有的校准技术，从而为机器学习领域带来价值。

此后，Facebook 计划将这些算法实施到 PyTorch 中，这将为已经使用其库的任何人提供快速、易于使用、强大的工具，并将显着提高校准技术的可见性和采用率。 PyTorch 是撰写本文时最受欢迎的机器学习库之一，这意味着该研究项目的输出将产生广泛而广泛的影响。

由于这个项目的实施，我们最终将拥有提供我们可以真正信任的不确定性的模型，因此当前的狂野西部机器学习方法将进一步推动与现实世界中的技术建立更安全的关系。

参考

- ¹ K. Suzuki, “计算机辅助诊断胸部和结肠 CT 中的机器学习:一项调查”,IEICE 信息和系统交易,第一卷。 E96-D,第 772–783 页,2013 年 4 月,24174708[pmid]。
- ² K. Murphy、B. van Ginneken、A. Schilham、B. de Hoop、H. Gietema 和 M. Prokop, “大规模使用局部图像特征和 k 近邻分类评估胸部 ct 中自动肺结节检测,”医学图像分析,卷。 13,没有。 5,第 757–770 页,2009 年。包括特别第 12 届国际医学影像和计算机辅助干预会议部分。
- ³ P. Koopman 和 M. Wagner, “自动驾驶汽车测试和验证的挑战”,SAE 国际运输安全杂志,第一卷。 4,第 15-24 页,2016 年 4 月。
- ⁴ B. Spanfelner、D. Richter、S. Ebel、U. Wilhelm 和 C. Patz, “将 ISO 26262 应用于驾驶员辅助系统”,2012 年。
- ⁵ M. Gharib、P. Lollini、M. Botta、E. Amparore、S. Donatelli 和 A. Bondavalli, “关于结合基于机器学习的组件的汽车系统的安全性:立场文件”,2018 年 6 月。
- ⁶ E. Sparks, “2021 年人工智能和机器学习的四个预测”。 HTTPS://www.forbes.com/sites/evansparks/2021/02/10/four-predictions-for 人工智能和机器学习-in-2021/, 2021。
访问:
2021-03-02。
- ⁷ M. Kull、M. Perelló-Nieto、M. Kängsepp、T. de Menezes e Silva Filho、H. Song 和 PA Flach, “超越温度标定:通过狄利克雷校准获得校准良好的多类概率,”CoRR,
卷。绝对/1910.12656, 2019。
- ⁸ A. Niculescu-Mizil 和 R. Caruana, “用监督学习预测好的概率”,第 625 页–632, 01 2005。
- ⁹ “优步的自动驾驶运营商因致命车祸被起诉。” https://www.bbc.co.uk/news/
技术-54175359,2020。访问时间:2021-03-02。
- ¹⁰ C. Rudin, “停止为高风险决策解释黑盒机器学习模型,而改用可预测的模型”,《自然机器智能》,卷。 1,第 206-215 页,2019 年 5 月。
- ¹¹ R. Metz, “人工智能如何在过去十年中统治我们的生活。” https://edition.cnn.com/2019/12/21/tech/artificial-intelligence-decade/,2019 年。访问时间:2021-03-02。
- ¹² Z. Ghahramani, “概率机器学习和人工智能”,《自然》,第一卷。 521,第 452-459 页,
2015 年 5 月。关于概率模型。
- ¹³ F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cournapeau、M. Brucher、M. Perrot 和 E. Duchesnay, “Scikit-learn:Python 中的机器学习”,机器学习研究杂志,第一卷。 12,第 2825-2830 页,2011 年。
- ¹⁴ “Scikit-learn 用户指南:1.16:概率校准。” https://scikit-learn.org/stable/
模块/校准.html。访问时间:2021-03-03。
- ¹⁵ M. Kull、TS Filho 和 P. Flach, “Beta 校准:有充分根据且易于实施的改进
关于二元分类器的逻辑校准”,第 20 届人工智能与统计国际会议论文集 (A. Singh 和 J. Zhu, 编辑),卷。 54 机器学习论文集
研究,(美国佛罗里达州劳德代尔堡),第 623-631 页,PMLR,2017 年 4 月 20-22 日。

¹⁶ A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang、J. Bai 和 S. Chintala, “Pytorch:一种命令式风格的高性能深度学习库，”
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 编辑) ,第一卷。 32,Curran Associates, Inc.,2019 年。

¹⁷ C. Guo、G. Pleiss、Y. Sun 和 KQ Weinberger, “关于现代神经网络的校准” ,CoRR,
卷。绝对/1706.04599, 2017。

¹⁸ TW Keelin, “Metalog 分布” ,决策分析,第一卷。 13,没有。 4,第 243-277 页,2016 年。

¹⁹ “2020 年机器学习和数据科学现状” 。 <https://www.kaggle.com/kaggle> 调查-2020。