机器学习分类算法的公平性审计

本文探讨了机器学习算法的公平性及其在社会关键领域的重要性，指出不公平的算法可能引发歧视和社会矛盾，同时加强隐私保护须减少对不必要数据的收集。基于假设检验，利用“马丁格尔随机过程”提出了一种公平性审计算法，并在信用卡数据集上验证。实验表明，该方法相比基线方法能更有效地降低假阳性率，为算法公平性审计提供有效手段。

文丨**张凡龙 钱钢** 南京审计大学计算机学院

算法审计是确保机器学习和人工智能系统以公平、合法、道德、安全的方式运作的关键手段，能够有效防止算法滥用，保护个人和社会的利益。本研究聚焦于分类算法的公平性审计方法。分类算法是机器学习中一种重要的监督学习方法，旨在根据输入特征将数据点分配到不同的类别或标签中。分类算法广泛应用于各种领域，能够帮助解决许多现实世界中的问题，其公平性日益受到人们关注。

机器学习算法公平性的重要性

机器学习算法在当今社会的多个重要领域中扮演着关键的角色。无论是招聘、贷款审批，还是司法决策，这些算法都在极大程度上影响着个体的生活和社会的运作。然而，随着机器学习技术的广泛应用，算法中的公平性问题也日益显现，并且其潜在的不公正性对社会的影响不容忽视。例如，在招聘过程中，算法能够通过分析历史数据（如简历、面试评分、员工绩效等）来筛选出最合适的候选人；在贷款审批中，算法通过分析借款人的财务状况、信用记录等信息来判断贷款风险；在司法领域，算法可能会被用于预测再犯风险、判决建议。

然而，如果算法在设计或训练过程中

未能充分考虑到公平性，就可能导致算法在决策中存在偏见，进而影响到某些群体。

例如，招聘算法可能会不公平地排除女性或少数族裔候选人，贷款审批算法可能会基于族群背景对某些人群进行不合理的风险评估，而司法算法可能会根据历史数据中的种族偏见，错误地预测某些族群的犯罪风险。这些问题不仅损害了被不公平对待群体的权益，还会造成更广泛的社会不满和不信任。

当公众意识到某些机器学习算法的决策存在偏见时，可能会对相关机构和社会制度产生严重的不信任。这种不信任感的积累，不仅会影响个体和社会之间的关系，也可能在更大的社会层面上引发矛盾和冲突。例如，如果某个金融机构的贷款算法被认为在种族或性别上存在偏见，受害群体可能会感到被排斥和不公正对待，进而对整个金融系统产生怀疑，甚至可能影响到该机构的客户信任度和市场声誉。随着时间推移，社会的不平等和不信任感可能加剧，导致更广泛的社会动荡。此外，在公共资源分配、社会保障等领域，不公平的算法同样可能引发民众的不满。例如，如果某个社会保障或福利发放系统的算法决策存在性别、年龄或地区上的偏见，可能会导致某些群体在资源分配上遭受不公，从而激化社会不平等，甚至可能引发抗议和社会运动。

在这种背景下，推动机器学习算法的公平性不仅是技术层面的需求，更是社会层面的必要性。越来越多的国家和地区出台了相关法律法规，对机器学习算法的公平性提出了明确要求。例如，欧盟的《通用数据保护条例》（GDPR）中就涉及对算法决策中个人数据保护和公平性的规定。企业和机构在使用机器学习算法时，必须确保其符合相关法律要求，否则将面临法律风险和处罚。

算法公平性对个人数据保护的影响

公平的算法决策减少不必要的数据收集：公平的机器学习算法在决策过程中会基于合理的因素和准确的数据进行判断，从而减少对不必要的个人数据的收集需求。这有助于降低个人数据被过度收集和滥用的风险，更好地保护个人隐私。例如，一个公平的信贷审批算法可能只需要关注申请人的收入、信用记录等与信贷风险直接相关的信息，而无须收集过多的个人敏感信息，如宗教信仰、政治观点等。

不公平的算法决策可能导致数据滥用，也可能会导致对某些群体的不合理关注和数据收集。例如，一个存在性别歧视的招聘算法可能会过度收集女性求职者的个人数据，试图寻找所谓的“女性特质”来支持其不公平的决策，这不仅侵犯了女性求职者的隐私，也加剧了性别不平等。

机器学习算法的公平性与个人数据保护是相辅相成、不可分割的。只有在保障个人数据安全和隐私的基础上，实现机器学习算法的公平性，才能充分发挥机器学习技术的优势，推动社会的公平、和谐与进步。

算法审计

算法审计作为一个相对较新的领域，尚未在学术界和实践中形成统一的定义和框架，不同的文献和研究为其提供了多种视角和定义。有文献指出，算法审计是一种通过反复且系统性地向算法输入数据，并观察相应输出，以便对其不透明的内部运作进行推断的方法。也有文献认为，算法审计是对算法的合法性、道德性以及安全性进行评估、缓解（风险）并提供保障的研究与实践活动。而其他文献也指出，算法审计的任务之一是针对存在偏见、歧视性或其他有害行为的算法系统进行审计。

机器学习算法在众多领域都有广泛的应用，但同时也引发了一系列公平性问题。

一个公平的社会应该是所有成员都能得到公正对待的社会，而偏见和歧视的存在则会破坏这种社会公正，引发社会矛盾和冲突。本文研究机器学习算法中的分类算法的公平性审计。

分类算法的公平性审计算法

假设检验是一种统计方法，用于评估样本数据是否支持某个特定假设。可以将算法是否公平分为以下两个假设：

零假设（H0）：算法是公平的。

备择假设（H1）：算法是不公平的。

于是，可以将审计算法公平性的问题表述为一个假设检验问题。其基本思想是：假设有人对机器学习算法的公平性表示怀疑。此人设计了一个迭代游戏，在审计过程中对结果进行下注。下注的结果是如果算法不公平，预期收益将很高；如果算法公平，则预期收益较小。因此，如果财富随时间增长，就可以认为算法不公平，即如果财富超过预定的阈值，则拒绝零假设，认为算法不公平。这一思想在数学上有严格的理论支撑，即马丁格尔（martingale）随机过程。其本质是：在给定过去信息的条件下，未来的期望值等于当前的值。这意味着马丁格尔随机过程在长时间内没有“趋势”，即没有预期的增长或下降。这被广泛用于需要建模不确定性的领域。

本文按照以上思想对机器学习中常见的分类算法进行公平性假设检验。分类算法被广泛应用于各种决策场景，如预测客户是否会违约、判断病人是否患有某种疾病、判断求职者是否符合某职位的要求等。评估的分类算法包括：Adaptive Boosting 算法、Decision Tree 算法、GaussianNB 算法、Gradient Boosting 算法、KNN 算法、Quadratic Discriminant Analysis 算法、Random Forest 算法、Ridge Classifier 算法。

这些算法的基本思想如下：AdaBoost 是一种集成学习方法，通过将多个弱分类器组合成一个强分类器来提高分类准确性。每个弱分类器在训练时都会特别关注前一次分类错误的样本。Decision Tree 算法是一种基于树结构的分类方法，通过不断地对特征进行划分来进行决策。Gaussian Naive Bayes (GaussianNB) 算法是一种基于贝叶斯定理的分类算法，假设特征之间条件独立，并且每个特征的分布服从高斯分布。

表 1 某银行消费者购买行为预测模型性能比较表

属性	属性说明
X0	信用卡客户 ID 号
X1	信用卡限额
X2	性别 (1 代表男性，2 代表女性)
X3	受教育程度 (1= 研究生 , 2= 大学 , 3= 高中 , 4= 其他 5= 未知 , 6= 未知)
X4	婚姻状况 (1= 已婚，2= 单身，3= 其他)
X5	年龄
X6-X11	过去六个月的还款情况，-1 代表按时还款；1 代表延时一个月还款；2 代表延时两个月还款……依次类推
X12-X17	过去六个月的账单数额情况。
X18-X23	过去六个月的还款数额情况
Y	目标属性，客户下个月还款违约情况 (1= 逾期，0= 未逾期)

来源：南京审计大学计算机学院

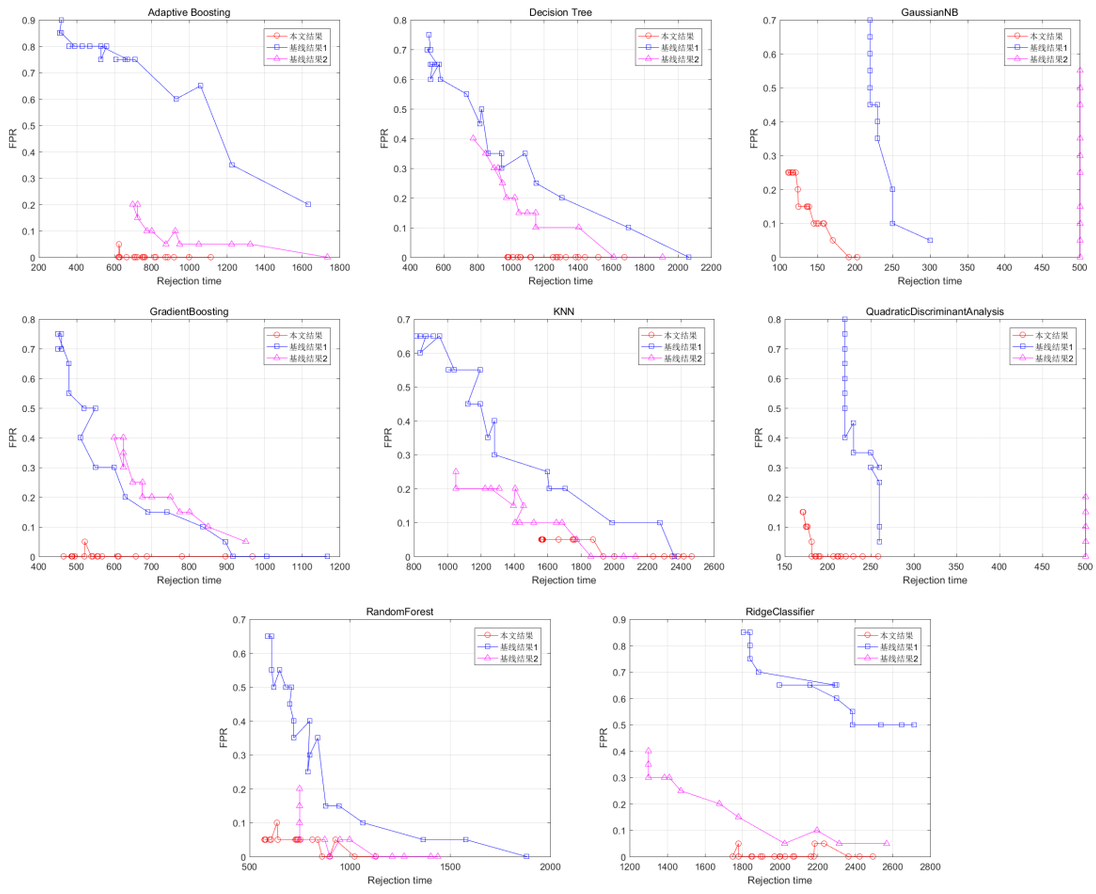
Gradient Boosting 算法是另一种集成学习方法，他通过多次训练弱学习器（通常是决策树），并在每次训练时最小化上一次模型的残差，从而逐步提高预测准确率。K-Nearest Neighbors (KNN) 算法是一种基于实例的学习方法，通过计算测试点与训练数据集中所有样本的距离，并选择最近的 K 个邻居进行投票来确定分类。Quadratic Discriminant Analysis(QDA) 算法是一种基于贝叶斯定理的分类方法，假设各类数据的特征遵循高斯分布，但与线性判别分析（LDA）不同，他为每个类别计算不同的协方差矩阵。Random Forest 算法是一种集成学习方法，通过训练多个决策树并在预测时对所有树的结果进行投票，来提高分类准确性。Ridge Classifier 算法是基于岭回归的分类方法，使用 L2 正则化来减少模型的复杂度，避免过拟合，尤其在特征之间存在高度共线性的情况下表现较好。

评估数据集是 Kaggle 提供的 default of credit clients 数据集，包括某银行（现金及信用卡发行）的信用卡持卡人的个人信息和用户信用卡消费还款信息。该数据集共有 3 万条数据和 25 个属性，具体信息如表 1 所示。

表中某些特征（如年龄、性别、婚姻状况等）可能在模型训练过程中扮演着重要角色。如果模型过度依赖这些敏感特征，可能会导致不公平的决策结果。在公平性检验中，检查各特征对预测结果的贡献以及是否导致了对某些群体的偏见是至关重要的。比较了多种机器学习模型在假阳性率 (FPR) 和拒绝时间 (Rejection Time) 下的表现。分析基线结果（不同 k 值的表现）和本文审计方法在各模型上的优劣。从图中结果可以得到如下结论：

基线结果的表现 （蓝线对应 k=200，粉线对应 k=500)：基线方法中，k 越小，审计的频率越高，因此拒绝时间较短，但假阳性率明显较高，表现不稳定。k 增加到 500（粉线）时，FPR 降低，但拒绝时间显著增加，效率降低。

本文审计方法的表现 （红线)：相比基线结果，表现出了更低的假阳性率，并且随着拒绝时间的增加能够更好地控制误差。对大部分模型，本文方法的假阳性率始终低于基线方法，并趋近于零。在 Quadratic Discriminant Analysis 和 GaussianNB 上，本文方法显著优于基线方法，假阳性率在很短的拒绝时间内快速下降。



来源：南京审计大学计算机学院

‖ 图 1 假阳性率（False Positive Rate，FPR）和拒绝时间在不同方法下的表现。X 轴：拒绝时间，Y 轴：假阳性率。比较了三种算法的审计结果，其中“基线结果”是指对于某个预先指定的自然数 k（即 k 表示每次采样的审计数量），一直等到我们收集到 k 次审计后，再进行检验。如果该检验没有拒绝（原假设），那就再收集另外的 k 次审计，然后重复上述操作。基线结果 1 和基线结果 2 分别对应 k=200 和 k=500。

结束语

算法公平性审计未来将越来越成为社会、法律和技术之间的交汇点。随着技术的不断演进、法律和伦理的逐步完善，算法公平性审计将变得更加全面、系统和高效。从自动化审计工具到跨学科的合作，再到公平性与可解释性的结合，算法审计的未来无疑会迎来更加复杂和精细化的发展。最终，算法公平性审计将成为推动社会公正、技术创新与伦理平衡的重要工具，确保人工智能技术能够造福全社会，而不偏向某一群体或利益。

责任编辑：徐培炎 xupy@staff.ccidnet.com