中国科学: 信息科学 2020 年 第 50 卷 第 8 期: 1255–1266
SCIENTIA SINICA Informationis
论文

公平性机器学习中基于分类间隔的歧视样本发现和消除算法

石鑫盛[1,2], 李云[1,2*]

1. 南京邮电大学计算机学院, 南京 210023
2. 南京邮电大学江苏省大数据安全与智能处理重点实验室, 南京 210023
* 通信作者. E-mail: liyun@njupt.edu.cn

收稿日期: 2019–05–30; 修回日期: 2019–08–24; 接受日期: 2019–12–05; 网络出版日期: 2020–08–05

国家自然科学基金 (批准号: 61603197, 61772284, 61876091, 61802205) 资助项目

摘要 公平性学习是机器学习领域的研究热点, 预防歧视的目的在于执行预测任务之前消除不公平训练集对于分类器的影响. 为了保证分类公平性和准确性, 本文通过发现和消除原始数据集中的歧视样本寻找生成公平数据集的方法, 即提出了一种基于分类间隔的加权方法用于处理二分类任务中的歧视现象, 并在 demographic parity 和 equalized odds 公平性判定准则上实现分类公平. 为了不影响分类准确性, 本文基于最大间隔原理将样本投影之后选出目标集, 对于目标集中的每个样本, 通过加权距离度量方法判定该样本是否具有歧视性, 并进行修正. 通过在 3 个真实数据集上与已有方法进行实验对比, 本文的方法能够获得更好的分类公平性和准确性, 并且不局限于特定的公平性判定准则和分类器.

关键词 公平性学习, 分类间隔, 目标集, 加权距离度量, 歧视性

1 引言

近年来, 机器学习越来越受到各界人士的关注. 然而, 机器学习系统容易受到历史数据的影响, 并对少数群体、弱势群体和历史上处于不利地位的群体产生歧视行为. 因此, 有必要使用公平性准则来约束机器学习系统在某些应用场景中的行为, 例如贷款、就业、刑事司法和广告, 并期望保护弱势群体, 从而在分类公平性和准确性之间达到一个平衡.

目前, 机器学习公平性的研究成果大致归为以下 3 大类.

第 1 类方法通过数据预处理消除歧视样本. 主要策略是通过修改训练集来平衡分类公平性和准确性, 如修改样本属性 [1,2]、修改样本标签 [3,4]、修改训练集大小 [5~7].

第 2 类方法在模型的训练过程中消除歧视. 主要策略是在目标函数中增加新的公平性约束条件、加入正则化项或者通过属性映射方法将原始属性向量映射为新的公平表示 [8~11]. 例如, Agarwal

引用格式: 石鑫盛, 李云. 公平性机器学习中基于分类间隔的歧视样本发现和消除算法. 中国科学: 信息科学, 2020, 50: 1255–1266, doi: 10.1360/SSI-2019-0112
Shi X S, Li Y. Discriminatory sample identifying and removing algorithms based on margin in fairness machine learning (in Chinese). Sci Sin Inform, 2020, 50: 1255–1266, doi: 10.1360/SSI-2019-0112

等 [8] 和 Dwork 等 [9] 在目标函数中加入公平性约束条件实现了分类公平性和准确性的双重目标. Lemoine 等 [10] 和 Madras 等 [11] 利用 GAN 网络 [12] 研究属性之间相关性, 并调整损失函数的梯度下降方向来获得分类公平.

第 3 类方法通过模型后处理消除歧视. 主要策略是先预训练一个分类器, 然后在考虑分类公平性的情况下移动分类器模型的决策边界 [13,14]. 例如 Fish 等 [13] 通过改变决策树的分割准则和剪枝策略使得修改后的决策树模型能够实现分类公平. Liu 等 [14] 在此基础上提出历史数据集 (history dataset) 会带来长期公平性问题, 并探索了静态公平性准则的长期影响.

本文基于数据预处理的方法, 主要通过修改歧视样本的标签来平衡分类公平性和准确性, 研究在预定义敏感属性前提下带有公平性约束的二分类任务. 这里的敏感属性指: 性别、种族、国籍、婚姻状况或年龄等等.

本文的方法基于聚类假设 “相似的样本应该获得相似的预测结果”, 并借鉴了 Luong 等 [3] 和 Zhang 等 [4] 的观点. 首先, 他们基于敏感属性将训练集划分为两个集合; 接着对于训练集中的每个样本通过相应的距离度量方法分别从这两个集合中选取出离该样本最近的 k 个样本, 当它们的标签差别很大, 即由于敏感属性不同而导致相似的样本无法获得相似的预测结果, 则可以认定这类样本为歧视样本; 最后, 修改歧视样本的标签, 利用修正后的训练集训练分类器模型, 并在未修正的原始测试集上测试模型的分类公平性和预测准确度.

Luong 等 [3] 采用归一化的曼哈顿 (Manhattan) 距离和重叠测量来计算样本间的距离并寻找 k 近邻样本. 他们的方法在特定场景下实验效果较好, 但是存在一些不足: (1) 使用所有属性进行距离计算, 没有考虑到不同属性对于歧视样本的发现重要性不同; (2) 随机选取 10% 的样本进行歧视发现和消除, 模型准确度损失过大. Zhang 等 [4] 作出了改进, 不再使用所有属性进行距离计算, 而是利用贝叶斯 (Bayes) 网络和有向分离原理 (directed separate) 将所有属性间的关系进行定向分离, 以获得对标签预测有重要影响的属性. 但是也存在一些不足: (1) 在计算样本之间的距离时, 该方法删除了对标签预测没有直接影响的属性, 但由于属性之间的相关性, 保留下来的属性可能会隐含那些被删除属性的信息; (2) 只使用对标签预测有重要影响的属性训练分类器, 并且将属性进行二值化处理, 这势必会影响到最终的分类准确度.

为了解决上述方法存在的问题, 本文提出了基于分类间隔的歧视样本发现和消除算法, 该方法既考虑对标签预测有重要影响的属性, 又关注其他可能会对标签预测产生潜在影响的属性. 首先, 为了确保歧视样本选取的合理性, 采用了一种基于间隔 (margin) [15] 的属性加权方法, 该方法的损失函数惩罚那些 k 近邻中与目标样本标签不同且距离较近的样本和与目标样本标签相同且距离较远的样本; 接着, 将学习到的权重用于计算样本间的距离和寻找 k 近邻样本; 并且, 为了保证分类公平性的同时尽可能降低模型精度的损失, 本文基于最大间隔原理将样本进行投影后选取目标集; 最后只在目标集中选取歧视样本并进行修正.

本文安排如下, 第 2 节介绍了公平性的定义准则和加权距离度量方法. 第 3 节提出了用于选取目标集的样本投影方法, 并在目标集中进行歧视样本的发现和消除. 第 4 节进行了对比实验. 第 5 节对全文进行了总结.

2 相关定义

在公平性机器学习中, 主要是二分类任务. 训练样例集表示为 (X, A, Y), 其中 X 表示不包含敏感属性的样本集. A 表示预定义的某一敏感属性, 例如种族、性别、年龄、婚姻状况等. Y 表示标

签. 本文假设 A 和 Y 都为二值变量. 特别地, 使用 (x_i, a_i, y_i) 表示一个带有标签的完整样本, 其中 $x_i = (x_i^1, x_i^2, \ldots, x_i^n) \in \mathbb{R}^n$ 表示该样本中除敏感属性以外的 n 维属性, 敏感属性 $a_i \in \{a_i^+, a_i^-\}$. 样本标签分为两类 $y_i \in \{y_i^+, y_i^-\}$, 其中 y_i^+ 表示正类, y_i^- 表示负类.

2.1 公平定义

机器学习的公平性定义准则主要有两种 —— demographic parity [9] 和 equalized odds [16].

定义1 (Demographic parity) 基于数据分布 (X, A, Y) 训练出的分类器 h, 如果预测结果 $h(X)$ 能够独立于敏感属性 A, 即 $\Pr\{h(X) = y^+ | A = a^+\} = \Pr\{h(X) = y^+ | A = a^-\}$, 则可以认为分类器 h 满足 demographic parity. 表示为如下不等式:

$$|\Pr\{h(X) = y^+ | A = a^+\} - \Pr\{h(X) = y^+ | A = a^-\}| \leqslant \tau, \tag{1}$$

其中阈值 τ 用作公平性约束.

定义2 (Equalized odds) 基于数据分布 (X, A, Y) 训练出的分类器 h, 如果预测结果 $h(X)$ 能够在给定标签 Y 的情况下条件独立于敏感属性 A, 即 $\Pr\{h(X) = y^+ | A = a^+, Y = y\} = \Pr\{h(X) = y^+ | A = a^-, Y = y\}$, $y \in \{y^+, y^-\}$, 则可以认为分类器 h 满足 equalized odds. 表示为如下不等式:

$$|\Pr\{h(X) = y^+ | A = a^+, Y = y\} - \Pr\{h(X) = y^+ | A = a^-, Y = y\}| \leqslant \tau, y \in \{y^+, y^-\}, \tag{2}$$

其中阈值 τ 用作公平性约束.

2.2 加权距离度量

本小节主要介绍加权距离度量方法以寻找 k 近邻样本. 为了计算样本之间的距离, Luong 等 [3] 首先将样本进行维度分解. 接着计算样本之间在每一维度上的属性值差异. 最后, 样本间的距离为所有维度属性距离之和. 具体做法如下.

给定数据集 S 中的两个样本 x_i, x_j, 假设每个样本都有除敏感属性外的 n 维属性, x_i^f 表示样本 x_i 中第 f 维属性的值. 若第 f 维为连续属性, 则样本 x_i, x_j 在第 f 维上的距离为

$$\mathrm{iso}\left(x_i^f, x_j^f\right) = \frac{|x_i^f - x_j^f|}{g^f}, \tag{3}$$

其中 g^f 表示所有样本在第 f 维属性上取到的最大值与最小值之差, 用于对结果进行归一化处理. 若第 f 维为离散属性, 则样本 x_i, x_j 在第 f 维属性上的距离为

$$\mathrm{nom}\left(x_i^f, x_j^f\right) = \begin{cases} 0, & \text{if } x_i^f = x_j^f, \\ 1, & \text{otherwise.} \end{cases} \tag{4}$$

最终, 样本 x_i 与 x_j 之间的距离为所有维度属性距离之和:

$$d(x_i, x_j) = \sum_{f=1}^{n} \mathrm{FD}\left(x_i^f, x_j^f\right), \tag{5}$$

其中

$$\mathrm{FD}\left(x_i^f, x_j^f\right) = \begin{cases} \mathrm{iso}(x_i^f, x_j^f), & \text{如果 } x_i^f, x_j^f \text{ 是连续属性值,} \\ \mathrm{nom}(x_i^f, x_j^f), & \text{如果 } x_i^f, x_j^f \text{ 是离散属性值.} \end{cases} \tag{6}$$

上述距离度量方法存在一定不足. 式 (5) 使用所有属性计算两个样本之间的距离, 但属性对于标签预测的重要性程度不同. 为了解决该问题, 本文使用加权距离度量方法, 定义如下.

定义3 给定数据集 S 中的两个样本 x_i, x_j, 它们之间的距离表示为

$$d\left(x_i, x_j\right)=\sum_{f=1}^{n} w^f \cdot \mathrm{FD}\left(x_i^f, x_j^f\right), \tag{7}$$

其中 $w^f \in [0,1]$ 表示第 f 维属性的权重. 本文采用了一种基于 Lmba [15] 的属性加权方法来获得每一维属性的权重 w^f, 具体是使用最大分类间隔原理惩罚那些目标样本的 k 近邻中标签与其不同的较近样本和标签相同的较远样本, 并通过梯度下降算法迭代求解损失函数. 本文通过该方法获得属性的权重向量 w, 然后将加权后的属性用于选定目标集 (见 3.1 小节) 和寻找歧视样本 (见 3.2 小节).

定义4 给定训练集 S, 样本 x_i, 属性的权重向量 $w=(w^1, w^2, \ldots, w^n)$. 则 x_i 的损失函数定义为

$$\begin{aligned} L_s\left(w, x_i\right) &= \sum_j t_{ij}\left\|x_i-x_j\right\|_w^2 + c\sum_{jp} t_{ij}\left(1-b_{ip}\right) h_{jp}\left(w, x_i\right), \\ h_{jp}\left(w, x_i\right) &= \left[\theta_i + \left\|x_i-x_j\right\|_w^2 - \left\|x_i-x_p\right\|_w^2\right]_+, \end{aligned} \tag{8}$$

其中常量 $c>0$ 为平衡系数. $b_{ip} \in \{0,1\}$ 表示样本 x_i 和 x_p 的标签是否相同, 若相同则为 1, 否则为 0. $t_{ij} \in \{0,1\}$ 表示样本 x_i 和 x_j 是否互为近邻样本, 若近邻则为 1, 否则为 0.

值得注意的是, 损失函数的第 1 项惩罚那些与样本 x_i 具有相同标签并且是其 k 近邻中距离较远的样本, 而第 2 项惩罚那些与 x_i 具有不同标签并且是其 k 近邻中距离较近的样本. Lmba [15] 属性加权方法尤其关注与 x_i 标签不同的近邻样本, 因此在损失函数的第 2 项加入预定义的参数 θ_i:

$$\theta_i = \left|\left\|x_i-\mathrm{nearmiss}\left(x_i\right)\right\|^2 - \left\|x_i-\mathrm{nearhit}\left(x_i\right)\right\|^2\right|, \tag{9}$$

其中 $\mathrm{nearhit}(x_i)$ 和 $\mathrm{nearmiss}(x_i)$ 分别表示与样本 x 距离最近并且标签相同或相反的样本.

3 歧视发现与消除

公平性机器学习的一个重要目的在于消除由于偏见造成训练的分类器带有分类歧视. 本文的目标是发现训练集中的歧视现象, 并修正训练集中的歧视样本, 最后使用修正后的训练集训练分类器模型.

本文所提的算法模型如图 1 所示, 首先从训练集 S 中筛选出需要进行歧视性检验的目标集 D (见 3.1 小节), 以便更高效地寻找训练集中的歧视样本, 并减少由于修改样本标签而对分类器精度造成的损失; 接着, 基于预先给定的二值敏感属性 A 将目标集 D 划分为保护集 D^+ 和非保护集 D^-. 例如, 假设数据集中敏感属性 A 是性别, 则敏感属性值为男性的样本被分到保护集 D^+ 中, 敏感属性值为女性的样本被分到非保护集 D^- 中; 然后寻找两个集合中的歧视样本 (见 3.2 小节), 并且为了保证公平, 本文从保护集和非保护集中选取相同数量的歧视样本进行修正 (见 3.3 小节); 最后利用修正后的训练集训练分类器模型, 并在测试集上验证分类公平性和预测准确度.

3.1 选定目标集

公平性机器学习的一个基本要求是在保证分类公平性的同时, 尽量不损害预测准确度. 本文的做法基于最大间隔原理, 利用在投影空间中的样本分布特性选定目标集, 进而更高效地从目标集中选取歧视样本并进行修正. 本文通过计算样本在每一个维度上的间隔, 得到间隔向量属性空间, 定义如下.

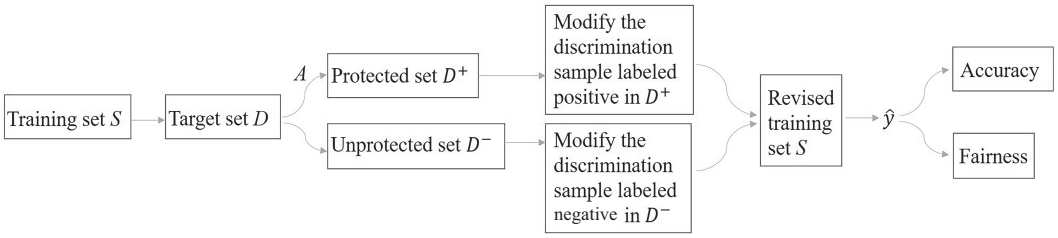

图 1 歧视样本发现和消除算法框架流程图

Figure 1 Framework diagram of discriminatory sample identifying and removing algorithm

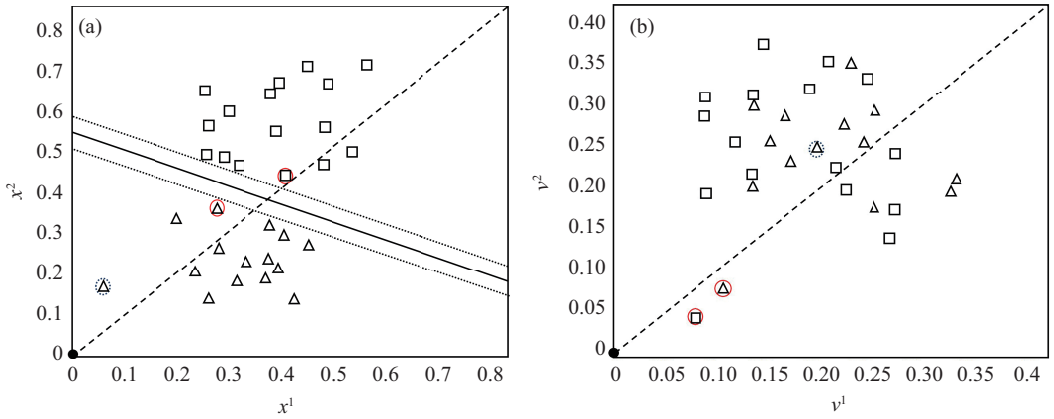

图 2 (网络版彩图) 对于属性投影后样本的分布特性变化的一个说明性实例. 原始属性空间 (a) 中的每一个样本沿各维度分解后投影到间隔向量属性空间 (b)

Figure 2 (Color online) An illustrative example of the change of distribution characteristics of samples after attribute projection. Each data point in the original feature space (a) is decomposed along each dimension and projected to the margin vector feature space (b)

定义5 给定数据集 S, 样本 $x_i = (x_i^1, x_i^2, \ldots, x_i^n)$. 对于样本 x_i, 在间隔向量属性空间中都存在一个对应的向量 $v_i = (v_i^1, v_i^2, \ldots, v_i^n)$, 如下所示:

$$v_i^f = \frac{1}{m}\sum_{l=1}^{m}\left(\left|\left\|x_i^f - \text{miss}\,(x_i)_l^f\right\| - \left\|x_i^f - \text{hit}\,(x_i)_l^f\right\|\right|\right), \tag{10}$$

其中 $\text{hit}(x_i)_l^f$ 和 $\text{miss}(x_i)_l^f$ 分别表示样本 x_i 的第 l 近邻样本的第 f 维属性值, 并且它们的标签分别与样本 x_i 标签相同或相反, m 表示样本 x_i 的近邻样本数量.

如式 (10) 所示, v_i^f 表示样本 x_i 与 m 个标签相同和相反的近邻样本在第 f 维属性上的间隔. 本文将各维度的间隔按照属性重要性程度不同进行加权求和, 属性权重向量 $w = (w^1, w^2, \ldots, w^n)$ (见 2.2 小节). 如下所示:

$$\text{sum}\,(x_i) = \sum_{f=1}^{n} w^f \cdot v_i^f, \tag{11}$$

若 $\text{sum}(x_i)$ 的值越大, 则表示样本 x_i 的不同类别近邻样本的差异性越大. 反之, 若 $\text{sum}(x_i)$ 的值越小, 则表示样本 x_i 的不同类别近邻样本的差异性越小.

如图 2 所示, 假设每个样本的属性维度为 2, 即 $x = (x^1, x^2)$. 图中的三角形和正方形分别代表不同标签的样本, 被圆圈圈出的 3 个样本在两个属性空间中的分布呈现出很大不同. 具体而言, 实心圆中

的两个样例在原空间中靠近所有样例的中心, 而在间隔向量属性空间中却成为离群点. 相反, 被虚线圈出的样例在原始属性空间中作为一个离群点出现, 但在变换后的空间中变得靠近中心. 这种现象产生的原因是由于间隔向量属性空间捕获的是不同类别近邻样本之间的差异性. 根据定义 5, 若 $\mathrm{sum}(x_i)$ 值越大, 即在间隔向量属性空间 (图 2(b)) 中离原点距离越远, 则样本 x_i 与标签不同的近邻样本距离越远, 与标签相同的近邻样本距离越近, 这符合 "相似的样本应该具有相同标签" 的聚类假设. 相反, 若 $\mathrm{sum}(x_i)$ 值越小, 即在间隔向量属性空间中离原点越近, 这表明样本 x_i 的近邻样本中存在很多标签不同的样本, 并且其中某些样本可能比同类近邻样本离 x_i 的距离还要近. 换言之, 若样本 x_i 为测试集中的样本, 则被分类器错误分类的可能性很大. 若样本 x_i 为训练集中的样本, 则该样本就是本文需要进行歧视性检验的目标样本. 因此, 本文将训练样本按照式 (10) 投影到间隔向量属性空间中, 并选取离原点最近的一些样本作为目标集.

3.2 歧视发现

本小节将叙述如何在选定的目标集中发现歧视样本. 给定训练集 S, 目标集 D, 样本标签 Y, 敏感属性 A. 首先, 基于敏感属性 A 将训练集 S 划分为两个集合. 具体而言, 敏感属性 A 取值为 a^+ 的样本被划分到集合 S^+ 中, 敏感属性 A 取值为 a^- 的样本被划分到集合 S^- 中. 接着对于目标集 D 中的任意样本 x_i, 基于 2.2 小节中提出的加权距离度量方法分别从集合 S^+, S^- 中挑选 k 近邻样本, 并计算它们的标签与 x_i 的标签相同的比例, 记为 p_1 和 p_2. 样本 x_i 是否为歧视样本可以通过 p_1 和 p_2 的差值判断, 即 $\mathrm{diff}(x_i, k) = p_1 - p_2$.

如算法 1 所示, 当其他属性相似, 仅敏感属性不同而导致近邻样本标签差异性很大, 则判定 x_i 为歧视样本. 这里的阈值 t 是由法律[1)2)3)4)]规定的, 例如, 英国法律对于性别歧视的容忍度阈值为 $t = 0.05$. 接下来在如下 3 个数据集中验证所提出的歧视性样本发现方法的有效性.

算法 1 歧视性样本发现算法

输入: 样本 $\{x_i, y_i\}$, 参数 t.

输出: 样本 $\{x_i, y_i\}$ 是否为歧视样本.

```
if y_i = y^- and diff(x_i, k) ⩽ −t then
    return True;
else if y_i = y^+ and diff(x_i, k) ⩾ t then
    return True;
else
    return False.
end if
```

- German Credit 数据集 [17] 包含 1000 个样本, 每个样本由 21 个属性描述. 这里的任务是预测顾客是否被允许借贷. 敏感属性是性别 (男性/女性), 正类标签为允许借贷.
- Adult Income 数据集[5)] 包含 48842 个实例, 每个样本由 14 个属性描述. 这里的任务是预测一个人年薪是否超过 50000. 敏感属性是种族 (白人/非白人), 正类标签为年薪超过 50000.

1) Australian Legislation. (a) Equal Opportunity Act Victoria State, 2010; (b) Anti-Discrimination Act Queensland State, 1991.

2) European Union Legislation. (a) Race Equality Directive, 2000; (b) Employment Equality Directive, 2000; (c) Equal Treatment of Persons, 2009.

3) U.K. Legislation. (a) Sex Discrimination Act, 1975; (b) Race Relation Act, 1976.

4) U.S. Federal Legislation. (a) Equal Credit Opportunity Act, 1974; (b) Fair Housing Act, 1968; (c) Employment Act, 1967; (d) Equal Pay Act, 1963; (e) Pregnancy Discrimination Act, 1978; (f) Civil Right Act, 1964, 1991.

5) https://github.com/frankhlchi/R-scorecard.

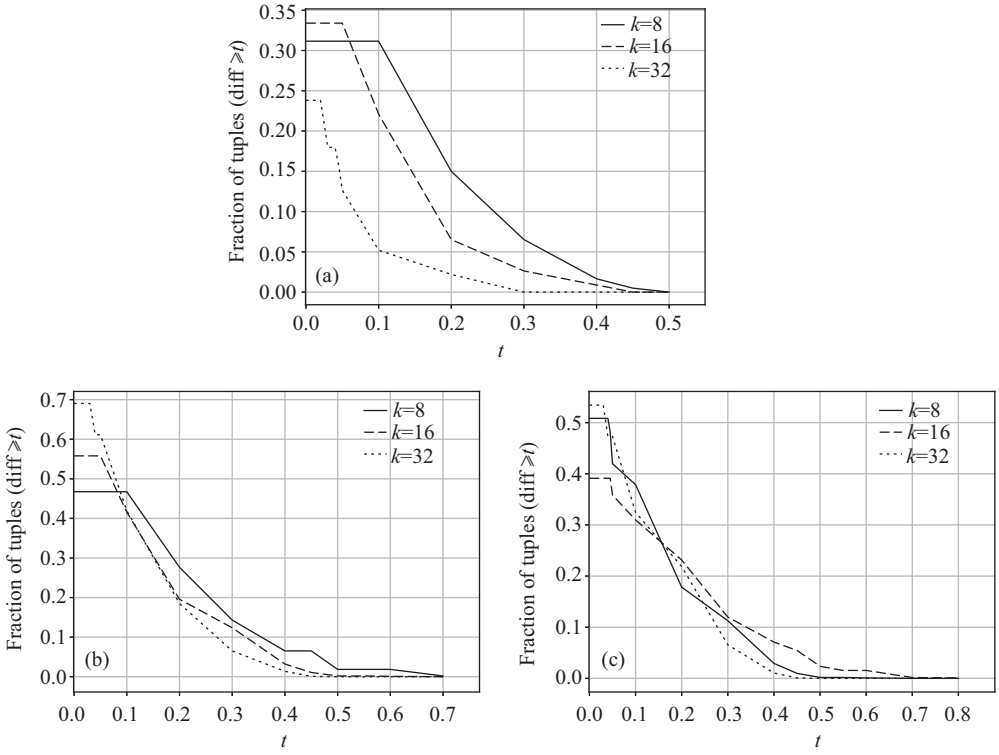

图 3 数据集 (a) German Credit, (b) Adult Income 和 (c) Dutch Census 中的歧视发现

Figure 3 Discrimination discovery on (a) German Credit, (b) Adult Income and (c) Dutch Census

• Dutch Census 数据集 [4] 包含 60421 个实例, 每个样本由 12 个属性描述. 这里的任务是预测一个人薪水是高薪还是低薪. 敏感属性是性别 (男性/女性), 正类标签为高薪.

例1 在数据集 German Credit [17] 中, 本文使用算法 1 测试不同阈值下 $\mathrm{diff}(x_i, k)$ 的变化情况, 并证明数据集 German Credit 中存在歧视现象. 图 3(a) 显示了数据集 German Credit 中选定不同参数 k 和阈值 t 条件下 $\mathrm{diff}(x_i, k)$ 的累积分布, 横坐标表示阈值 t, 纵坐标表示对应阈值条件下数据集 German Credit 中歧视样本的比例. 较小的阈值 t 意味着较小的歧视容忍度, 即更严格的公平性. 如图 3 所示, 当阈值 t 为 0.05 时, 超过 15% 的顾客的 $\mathrm{diff}(x_i, k) \geqslant 0.05$, 这意味着男性的可借贷率比女性的可借贷率更高. 特别地, 对于 $k = 16$ 或 32, 即使公平性约束降低到 0.1, 依然有超过 20% 的不公平率.

例2 对于数据集 Adult Income, 本文设置不同的阈值 t 和参数 k 来观察 $\mathrm{diff}(x_i, k)$ 的变化情况. 在图 3(b) 中, 歧视现象更为严重. 分布折线都显示超过 40% 的人有 $\mathrm{diff}(x_i, k) \geqslant 0.1$, 这意味着白人比非白人更容易获得超过 50000 的薪资.

例3 对于数据集 Dutch Census [4], 本文同样设置不同的阈值 t 和参数 k 来观察 $\mathrm{diff}(x_i, k)$ 的变化情况. 在图 3(c) 中, 可以看到对于女性的歧视性现象依然存在.

图 3 显示 3 个数据集都存在不同程度的歧视现象. 具体而言, 由聚类假设 "相似的样本应该具有相同标签" 可知, 若不考虑敏感属性差异, 由敏感属性划分的两个集合 $\{S^+, S^-\}$ 中的样本被分到正类或者负类的概率应该基本相等. 但实际情况是: 由于敏感属性的差异导致这两个集合中样本被分到正类或者负类的概率差别很大. 那么在实际生活中, 使用这些数据集训练出的分类器对于某些弱势群体

或者小众群体会造成歧视, 进而可能会引发社会问题.

3.3 歧视消除

上文证明了数据集 German Credit, Adult Income 和 Dutch Census 中都存在不同程度的歧视, 如果仍然使用这种带有偏见的数据集训练分类器模型, 势必会对未来的分类公平造成影响. 本文的主要歧视性样本消除策略是在训练分类器之前修改训练集中的样本标签, 使用修改后的训练集训练分类器模型, 从而保证分类的准确度和公平性. 歧视性样本消除策略如算法 2 所示.

算法 2 歧视性样本消除算法

输入: 训练集 $S=\{x_i,a_i,y_i\}_{i=1}^{n}$, 参数 t,k,z, 修改标签的数量 M.
输出: 公平的分类模型 h.

初始化 $R^+=R^-=\emptyset$, $j=0$.
基于 3.1 小节所介绍的方法从训练集中筛选出目标集 D.
将目标集 D 通过敏感属性 A 的值划分为保护集 D^+ 和非保护集 D^-.
for $i=1,2,\ldots,z$ **do**
    $R^+=R^+\cup\left\{\mathrm{diff}\,(x_i,k)>t,y_i=y^+,x_i\in D^+\right\}$;
    $R^-=R^-\cup\left\{\mathrm{diff}\,(x_i,k)<-t,y_i=y^-,x_i\in D^-\right\}$;
end for
while $j\leqslant M$ **do**
    从 R^+ 中随机挑选一个样本 x, 将标签从 y^+ 修改为 y^-;
    从 R^- 中随机挑选一个样本 x, 将标签从 y^- 修改为 y^+;
    j++;
end while
用 R^+, R^- 中修改标签的样本替换训练集 S 中的对应样本, 并用修正后的训练集训练分类模型 h.

在算法 2 中, 首先从训练集中筛选出需要进行歧视检验的目标集 (见 3.1 小节); 接着使用敏感属性将目标集划分为保护集和非保护集, 并基于聚类假设通过算法 1 寻找出目标集中的歧视样本; 然后随机选择其中的 M 个歧视样本, 并修正它们的标签; 最后用修正的训练集训练分类器模型.

其中 $z=\min\{|D^+|,|D^-|\}$, 需要修改标签的数量是 M, 定义如下:

$$M=\left\lceil n^{a^-}\cdot(\Delta\mathrm{Pr}-\tau)\right\rceil/2, \tag{12}$$

其中阈值 τ 表示学习模型的分类公平性约束, n^{a^-} 表示训练集中敏感属性值为 a^- 的样本的数量. 如果本文使用 demographic parity (见 2.1 小节) 作为公平性准则, 则 $\Delta\mathrm{Pr}=|\mathrm{Pr}\{h(X)=y^+|A=a^+\}-\mathrm{Pr}\{h(X)=y^+|A=a^-\}|$. 如果本文使用 equalized odds 作为公平性准则, 则 $\Delta\mathrm{Pr}=|\mathrm{Pr}\{h(X)=y^+|A=a^+,Y=y\}-\mathrm{Pr}\{h(X)=y^+|A=a^-,Y=y\}|$.

4 实验结果

本节检验 3.3 小节中提出的歧视性样本消除算法在真实数据集上的分类效果, 将训练好的学习模型在测试集上的预测准确率和公平性准则 demographic parity (Fairness-DP) 和 equalized odds (Fairness-EO) 作为度量标准. 本文使用 German Credit, Adult Income 和 Dutch Census 这 3 个数据集进行实验, 把未做任何处理的数据集训练的分类器作为 Baseline, 并与 Luong 等 [3], Zhang 等 [4] 提出的算法进行比较. 实验用到了 3 个分类器: Logistic Regression (log.reg.), AdaBoost 和 Support Vector Machine (SVM). 本文进行了 10 次实验, 去除最好和最坏情况, 并对结果取平均值.

表 1 Baseline, Luong 等的方法, Zhang 等的方法以及本文方法在 Dutch Census 数据集上的歧视性样本消除效果

Table 1 Baseline, Luong et al., Zhang et al. and our method of discriminating defense results on the Dutch Census

Classifier	Baseline			Luong et al.		
	Accuarcy	Fairness-DP	Fairness-EO	Accuarcy	Fairness-DP	Fairness-EO
log.reg.	0.8436	0.8897	0.8601	0.7893	0.9252	0.9108
AdaBoost	0.8562	0.9085	0.9004	0.8089	0.9271	0.9342
SVM	0.8487	0.8998	0.8680	0.8024	0.9233	0.9472
Classifier	Zhang et al.			Our method		
	Accuarcy	Fairness-DP	Fairness-EO	Accuarcy	Fairness-DP	Fairness-EO
log.reg.	0.7948	0.9730	0.9744	0.8388	0.9795	0.9894
AdaBoost	0.8378	0.9645	0.9624	0.8478	0.9806	0.9847
SVM	0.8096	0.9836	0.9731	0.8395	0.9851	0.9703

本文设定参数 $k \in \{8, 16, 32\}$, $t \in \{0.025, 0.05, 0.075, 0.1\}$, $m = 16$. 对于 German Credit 和 Dutch Census 数据集, 设定 $\tau = 0.05$ (式 (12)), 而对于 Adult Income 数据集, 设定 $\tau = 0.025$. 对于每个参数都用歧视性样本消除算法检验效果. 当 $k = 16, t = 0.05$ 时, 结果最稳定, 效果最好. 若 k 取值太大时, 相当于将很多不是太相似的样本也加入近邻样本中, 进而影响实验效果; 若 k 取值太小时, 结果波动太大. 而若 t 取值太大时, 本文能够修改的歧视样本数量太少, 效果不明显; 若 t 取值太小时, 歧视性约束太强, 造成分类准确度下降过多.

表 1 展示了 3 种方法以及不考虑歧视情况下的 Baseline 在数据集 Dutch Census 上的歧视性样本消除效果. 实验结果如图 4 所示, 横坐标代表分类准确性和公平性准则 demographic parity, equalized odds, 纵坐标代表这 3 个指标的具体数值. 图 4 和表 1 中的结果均表明通过数据预处理获得公平的学习模型在分类公平性增加的同时会造成不同程度预测准确性的降低. 但不同的是, 与 Baseline 相比, Zhang 等 [4] 和本文的方法能够获得很好的分类公平性, 而 Luong 等 [3] 提出的方法的分类公平性效果提升不明显. Luong 等 [3] 提出的方法由于没有考虑到不同属性对于发现歧视样本的重要性不同, 因此处理后的数据集中仍然有很大的歧视, 并且随机选取一部分样本进行歧视发现和消除, 模型准确度损失过大. Zhang 等 [4] 和本文的方法都能够实现分类公平. 但是如图 4 和表 1 所示, Zhang 等 [4] 只使用了对标签预测有重要影响的属性, 并进行二值化处理, 尽管能够确保公平性但预测准确度的损失较大. 本文使用基于最大间隔原理来寻找近邻样本, 使得歧视样本的选取更合理, 并且为了保证分类公平性的同时尽可能降低模型精度的损失, 基于最大间隔原理将样本进行投影后选取目标集, 只在目标集中选取歧视样本并进行修正. 总体而言, 与 Luong 等 [3], Zhang 等 [4] 和 Baseline 比较, 本文的方法在不损害太多预测准确性的情况下获得很好的分类公平性.

5 总结

本文提出了一种基于分类间隔的加权方法来解决由数据集造成的歧视问题. 基于聚类假设 “相似的样本应该获得相似的预测结果”, 通过属性加权更合理地选择近邻样本, 并基于最大间隔原理将样本投影后选定目标集的方法来更准确地寻找歧视样本, 从而降低修改标签带来的精度损失. 本文所提出的歧视性样本消除算法在多个分类器中都有很好的效果, 并且适用于多个公平性准则.

在训练期间给定敏感属性的前提下, 本文的方法在分类公平性和预测准确度之间获得了很好的平

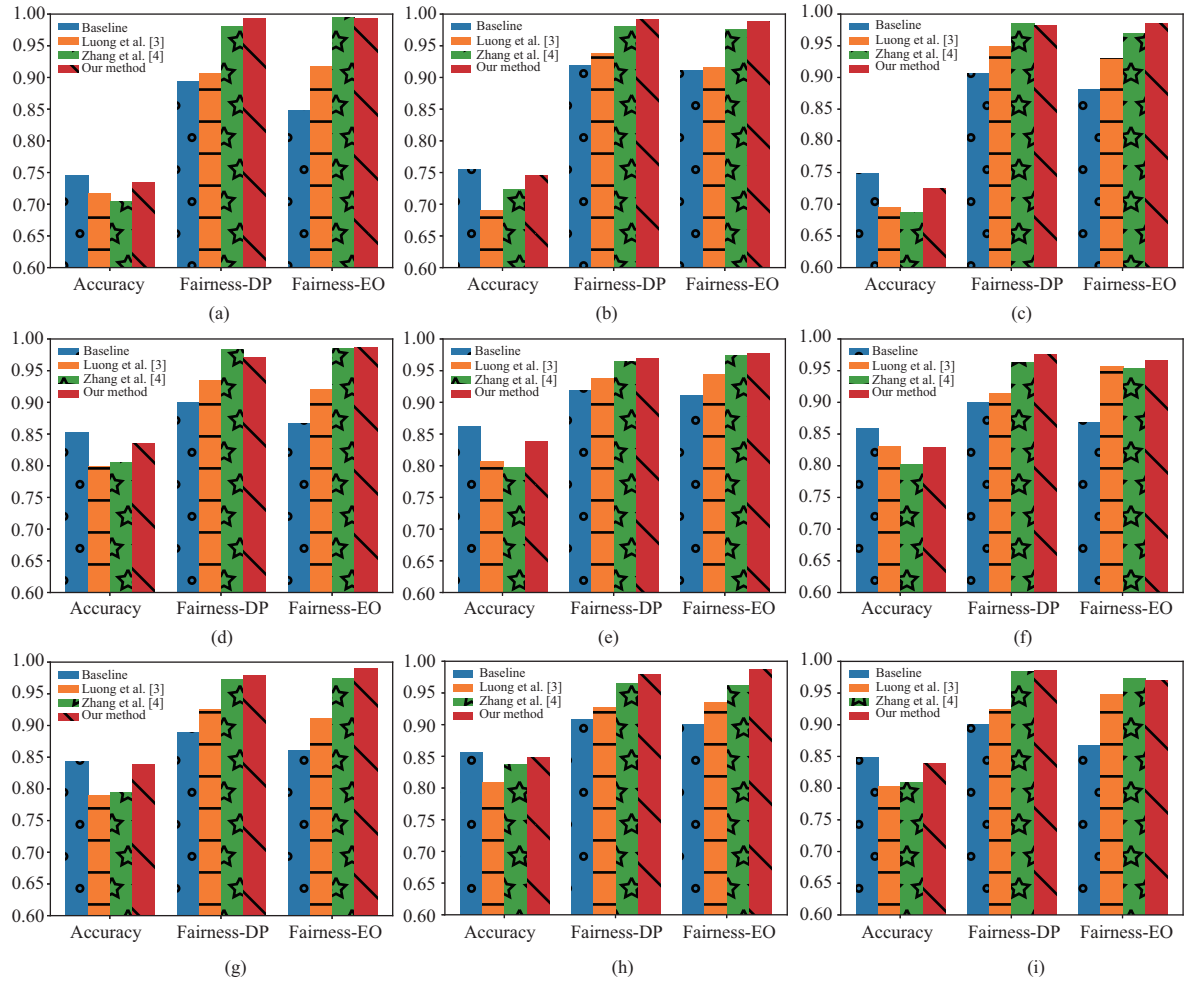

图 4 (网络版彩图) 修改训练集后在 3 个分类器上的预测准确度和分类公平性. (a)~(c), (d)~(f), (g)~(i) 分别表示在 German Credit, Adult Income 和 Dutch Census 数据集上使用 3 个分类器 (log.reg., AdaBoost, SVM) 的实验结果

Figure 4 (Color online)Accuracy and fairness on three classifiers after modifying datasets. (a)~(c), (d)~(f), (g)~(i) show the experimental results of using three classifiers (log.reg., AdaBoost and SVM) on German credit, Adult income and Dutch census datasets, respectively

衡. 但在某些场景中, 无法确定哪个属性或属性的组合是敏感的. 我们在后续工作中会更关注敏感属性的寻找.

参考文献

1 Feldman M, Friedler S A, Moeller J, et al. Certifying and removing disparate impact. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, 2015. 259–268

2 Zemel R, Wu Y, Swersky K, et al. Learning fair representations. In: Proceedings of the 35th International Conference on Machine Learning, 2018. 80: 3384–3393

3 Luong B T, Ruggieri S, Turini F. K-NN as an implementation of situation testing for discrimination discovery and prevention. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, 2011. 502–510

4 Zhang L, Wu Y, Wu X. Achieving non-discrimination in data release. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, 2017. 1335–1344
5 Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowl Inf Syst, 2012, 33: 1–33
6 Kamiran F, Karim A, Zhang X. Decision theory for discrimination-aware classification. In: Proceedings of 2012 IEEE 12th International Conference on Data Mining, Brussels, 2012. 924–929
7 Zliobaite I, Kamiran F, Calders T. Handling conditional discrimination. In: Proceedings of 2011 IEEE 11th International Conference on Data Mining, Vancouver, 2011. 992–1001
8 Agarwal A, Beygelzimer A, Dudik M, et al. A reductions approach to fair classification. In: Proceedings of the 35th International Conference on Machine Learning, 2018. 80: 60–69
9 Dwork C, Hardt M, Pitassi T, et al. Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, 2012. 214–226
10 Zhang B H, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, 2018. 335–340
11 Madras D, Creager E, Pitassi T, et al. Learning adversarially fair and transferable representations. In: Proceedings of the 35th International Conference on Machine Learning, 2018. 80: 3384–3393
12 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, Montreal, 2014. 2672–2680
13 Fish B, Kun J, Lelkes D. A confidence-based approach for balancing fairness and accuracy. In: Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, 2016. 144–152
14 Liu L T, Dean S, Rolf E, et al. Delayed impact of fair machine learning. In: Proceedings of the 35th International Conference on Machine Learning, 2018. 80: 3150–3158
15 Li Y, Lu B L. Feature selection based on loss-margin of nearest neighbor classification. Pattern Recogn, 2009, 42: 1914–1921
16 Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Proceedings of Advances in Neural Information Processing Systems, Barcelona, 2016. 3315–3323
17 Lichman M. UCI Machine Learning Repository, 2013. http://archive.ics.uci.edu/ml

Discriminatory sample identifying and removing algorithms based on margin in fairness machine learning

Xinsheng SHI[1,2] & Yun LI[1,2*]

1. *School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing* 210023*, China*;
2. *Jiangsu Key Laboratory of Big Data Security and Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing* 210023*, China*
* Corresponding author. E-mail: liyun@njupt.edu.cn

Abstract Fairness learning is one of research hotspots in machine learning. The purpose of preventing discrimination is to eliminate the impact of unfair training sets on classifiers before performing prediction tasks. To ensure the fairness and accuracy of classification, this paper presents a method for generating fair data sets by identifying and eliminating discriminatory samples in original data sets. This is a margin-based weighted method for dealing with discrimination in binary classification tasks and obtaining the demographic parity and equalized odds. To improve the classification accuracy, the target set is selected after projecting based on the margin principle. For each sample in the target set, a weighted distance measurement method is used to identify the discriminatory sample and then correct it. The experimental results on three real data sets demonstrate that the proposed method can obtain better classification fairness and accuracy than existing methods; the conclusion is not limited to specific fairness criteria or classifiers.

Keywords fairness learning, classification margin, target set, weighted distance metric, discriminatory

Xinsheng SHI was born in 1994. He received his Bachelor's degree from Guilin University of Electronic Technology. He is now a third-year graduate student at Nanjing University of Posts and Telecommunications. His current research direction is fair machine learning. He is a member of the Jiangsu Key Laboratory of Big Data Security and Intelligent Processing.

Yun LI was born in 1974. He received his Ph.D. degree in computer science from Chongqing University, Chongqing, China. He was a post-doctoral fellow in the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He is a professor in the College of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China. His current research interests include machine learning, data mining, and parallel computing.