计算机科学
COMPUTER SCIENCE

公平谱聚类方法用于提高簇的公平性

徐夏, 张晖, 杨春明, 李波, 赵旭剑

引用本文

徐夏, 张晖, 杨春明, 李波, 赵旭剑公平谱聚类方法用于提高簇的公平性[J]. 计算机科学, 2023, 50(2): 158-165.

XU Xia, ZHANG Hui, YANG Chunming, LI Bo, ZHAO Xujian. Fair Method for Spectral Clustering to Improve Intra-cluster Fairness [J]. Computer Science, 2023, 50(2): 158-165.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211100279

公平谱聚类方法用于提高簇的公平性

徐 夏 张 晖 杨春明 李 波 赵旭剑
西南科技大学计算机科学与技术学院 四川 绵阳 621010
(hsuhsia@qq.com)

摘 要 最近,算法的公平性问题引起了机器学习领域学者的广泛讨论。鉴于谱聚类在现代数据科学中的广泛流行,研究谱聚类的算法公平性是一个至关重要的话题。现有的公平谱聚类算法主要存在两个缺点:1)公平性能差;2)仅在单个敏感属性下工作。文中将公平问题视为一种约束谱聚类问题,通过求解约束谱聚类的可行解集,提出了一种非规范化公平谱聚类方法(Unnormalized Fair Spectral Clustering,UFSC),用于提升公平性能。此外,文中还提出了一种适用于多个敏感属性约束的公平聚类算法(Multi-sensitive Attributes Fair Spectral Clustering,MFSC)。在多个真实数据集上进行了实验,结果表明,UFSC和MFSC算法比现有的公平谱聚类算法生成的聚类结果更加公平。

关键词:算法公平性;公平谱聚类;约束谱聚类;机器学习;数据分析
中图法分类号 TP301

Fair Method for Spectral Clustering to Improve Intra-cluster Fairness

XU Xia,ZHANG Hui,YANG Chunming,LI Bo and ZHAO Xujian
School of Computer Science and Technology,Southwest University of Science and Technology,Mianyang,Sichuan 621010,China

Abstract Recently,the fairness of the algorithm has aroused extensive discussion in the machine learning community. Given the widespread popularity of spectral clustering in modern data science,studying the algorithm fairness of spectral clustering is a crucial topic. Existing fair spectral clustering algorithms have two shortcomings:1) poor fairness performance;2) work only for single sensitive attribute. In this paper,the fair spectral clustering problem is regarded as a constrained spectral clustering problem. By solving the feasible solution set of constrained spectral clustering,an unnormalized fair spectral clustering(UFSC) method is proposed to improve fairness performance. In addition,the paper also proposes a fair clustering algorithm suitable for multiple sensitive attribute constraints. Experimental results on multiple real-world datasets demonstrate that the UFSC and MFSC are fairer than the existing fair spectral clustering algorithms.

Keywords Algorithm fairness,Fair spectral clustering,Constrained spectral clustering,Machine learning,Data analysis

1 引言

聚类分析是数据挖掘和机器学习中的一个基本问题,它试图将数据对象划分到不同的簇中,以最大化簇内的相似性,最小化簇间的相似性。聚类算法被广泛应用于自然语言处理[1]、图像分割[2]、基因表达分析[3]、模式识别[4]和数据挖掘[5]等领域。目前,已有多个版本的聚类算法被提出,例如划分聚类、密度聚类、层次聚类、谱聚类和模糊聚类。谱聚类是一种基于图论的聚类算法,它将数据对象看作无向图中的顶点,任意两个顶点之间的相似度作为连接两个顶点的边的权重。通过求解无向图最小的分割问题,将无向图切分成互不连通的子图,一个子图被视为一个簇,从而完成聚类任务。与其他聚类算法相比,谱聚类算法具有一些明显的优点:1)易于编程实现;2)可以有效地求解标准线性代数;3)聚类结果通常优于传统聚类算法,如 k-means。;4)它对数据分布有很强的适应性。因此,谱聚类在近年来已成为最流行的现代聚类算法之一。

公平机器学习始于Dwork等[6]的早期开创性工作,随着机器学习技术越来越多地应用于日常决策(如贷款审批和法院判决建议[7]等),公平机器学习逐渐成为机器学习领域的热门话题之一。研究界投入了大量的精力为分类和回归等监督学习任务开发公平算法[8-13]。最近,研究界将目光转向了无监督学习,特别是聚类任务。基于DI原则[9],Chierichetti等[14]首先提出了公平聚类的概念。该公平概念指出,每个保护组(保护组是某个特定属性的取值,这个特定属性在公平聚类任务中被称为敏感属性,它的一个取值被称为一个保护组,

到稿日期:2021-11-28 返修日期:2022-06-24
基金项目:四川省科技厅重点研发项目(2021YFG0031);四川省省级科研院所科技成果转化项目(2022JDZH0035)
This work was supported by the Key R&D Project of Science & Technology Department of Sichuan Province(2021YFG0031) and Scientific and Technological Achievements Transformation Project of Sichuan Provincial Scientific Research Institute(2022JDZH0035).
通信作者:张晖(zhanghui@swust.edu.cn)

如性别属性,它可能拥有两个组,即男性和女性)在每个簇中应该成比例的出现。这一定义催生了一系列关于公平聚类算法的研究[15-21]。尽管研究界为开发公平聚类算法做出了许多努力,但是大多数公平聚类的研究都集中在基于中心的聚类算法上,仅在 Kleindessner 等[18]的工作中考虑了谱聚类中的算法公平性问题。然而,Kleindessner 等的方法存在不能保证得到一个公平的结果以及在公平性指标上表现一般等问题。此外,Kleindessner 等的方法只适用于一个敏感属性,而在公平聚类中,数据对象经常处于多个敏感属性中。例如,一个非裔美国女性,她可能处于人种、国家、性别等多个敏感属性中。在本项研究中,我们探讨如何提升谱聚类在公平指标上的性能以及开发出适用于多个敏感属性的公平聚类算法。

为了解决上述问题,我们提出了一种非规范化的公平谱聚类方法和一种多敏感属性的公平谱聚类方法。具体来说,我们在谱聚类中引入了一个叫做公平上下界[14]的概念。公平上下界认为,如果保护组在每个簇中的比例处于设定好的上下界内,那么这个聚类被认为是公平的。我们展示了如何将公平上下界概念与谱聚类整合,形成公平约束的谱聚类问题,并给出了公平约束谱聚类问题的解决方法。此外,我们给出了一种适用于多个敏感属性的公平谱聚类解决方案。

2　相关工作

近年来,研究界在为机器学习算法提供公平性保证或研究公平变体方面做了大量的工作,许多相关的公平算法已经被提出用于监督学习。在本项研究中,我们关注无监督学习领域的公平性问题,重点是公平聚类问题。

Chierichetti 等[14]首先在聚类领域引入了算法公平的概念。在他们的工作中,数据集被预处理成多个“平衡”(即他们对公平的定义)的子集(称为 fairlet),然后使用 k-center 和 k-median 算法对预处理好的数据集进行聚类。由于他们的方法寻找 fairlet 代价很高,并且只给出了基于 k-center 和 k-median 的两种公平聚类算法,以及敏感属性要求是具有两个值的单个敏感属性等诸多限制条件,他们的方法并不是非常成功。但他们的工作为公平聚类这个领域的研究开拓了思路,尤其是他们首次定义了聚类中公平的概念。后来,一些后续工作扩展了他们的想法。Schmidt 等[22]将 fairlet 扩展为适用于多值单个敏感属性的 k-均值聚类对象。在这项工作中,他们建议在“coreset”上解决公平聚类问题,coreset 是原始数据集的一个代表性子集。通过求解 coreset 上的公平聚类问题,可以给出原问题的近似解。尽管 coreset 加速了寻找 fairlet 的过程,但它仍然需要至少二次方的运行时间。Backurs 等[17]建议将输入数据嵌入层次分明的树(Hierarchically well-Separated Tree)中,以加速 fairlet 分解。文献[19]提出了一个基于中心聚类的通用公平聚类框架,将文献[14]提出的公平概念扩展为保护组在原始集中所占比例的上下界。该框架分为两个步骤,首先利用普通聚类算法得到聚类中心,然后将数据对象公平地分配到上一步得到的聚类中心。其中,公平分配被看作一个线性规划问题。文献[20]的工作与前者类似,不同的是,为了防止保护组的过度表示,他们只给出了保护组的上限。Rösner 等[15]考虑了 k-center 聚类算法的多种公平变体,并开发了一种用于多值单个敏感属性的常数因子近似 k-center 聚类算法。Bercea 等[16]改进了文献[15]中的结果,并针对几种经典聚类目标给出了双准则常数因子近似算法。文献[23]研究了公平聚类的两种变体,一种是最小散度聚类,另一种是公平上界聚类。Kleindessner 等[24]提出了一种简单的公平 k-center 聚类算法。该方法将一个簇视为原始数据集的一个摘要,并为每个簇生成公平的摘要。Ziko 等[25]通过在目标函数中添加公平损失项来把公平约束整合到聚类过程中。他们将公平损失项定义为保护组在敏感属性的概率分布与数据集中概率分布之间的 KL 散度。然而,该方法是仅针对一个敏感属性而设计的。文献[26]考虑了比例质心聚类问题,并提出了一个独立的公平性概念。Jung 等[27]提出了个体公平性的概念,该概念要求每个对象在某种程度上邻近一个中心,其中“某种程度”的取值依赖于该对象的 k 个近邻。Davidson 等[28]的研究表明,线性规划方法可以用于计算任意两个保护组的公平聚类。最近,Ghadiri 等[29]和 Abbasi 等[30]分别独立地提出了社会公平聚类问题。Makarychev 等[31]改进并推广了文献[29-30]提出的方法。

Chhabra 等[32]研究了层次聚集聚类(Hierarchical Aggregative Clustering,HAC)中的公平性问题。他们通过启发式方法将公平性约束引入到贪婪的 HAC 算法中,并提出了一种时间复杂度为三次方的公平性方法和一种对 HAC 输出结果进行校正的公平 HAC 方法。Ahmadian 等[33]也考虑了层次聚类中的公平性问题。他们将文献[20]中扩展的 fairlet 引入到层次聚类中,并提供了寻找 fairlet 的强计算下界和一种多项式时间的近似算法。Quy 等[34]研究了教育领域的公平问题,并给出了两种公平的解决方案:层次公平聚类和基于划分的公平聚类。与本研究最相关的是 Kleindessner 等[18]的工作。在他们的工作中,通过对簇指标矩阵施加线性公平约束,实现了公平谱聚类方法。但是,该方法存在公平指标表现一般,且只适用于单个敏感属性等缺点。在我们的方法中,公平上下界[14]概念被整合到谱聚类过程中,以提升聚类结果的公平性。对于数据对象处于多个敏感属性的情况,我们也给出了一种解决方案。

3　背景知识

本节回顾了非规范化谱聚类算法并定义了非规范化的公平谱聚类问题,从而引入了相关术语并为后续的工作奠定了基础。

对于任意 $n\in\mathbb{N}$,令 $[n]=\{1,2,\cdots,n\}$,$\mathbf{I}_n$ 表示一个 $n\times n$ 的单位矩阵。令 $G(P,E)$ 表示一个无向图,其中顶点集合 $P=\{p_1,p_2,\cdots,p_n\}$ 拥有 n 个节点,$E=\{e_{ij}\mid i,j\in P\}$ 是节点 i 和节点 j 之间的边。令 a_{ij} 表示边 e_{ij} 的权重,它衡量了节点 i 和节点 j 之间的相似性,a_{ij} 满足 $a_{ij}>0$ 且 $a_{ij}=a_{ji}$。给定 $k\in\mathbb{N}^*$,令 C_j,$j\in[k]$ 分别表示图 G 子图中的节点集合。对于任意的 $g,h\in[k]$,C_g,C_h 之间的切图权重被定义为 $Cut(C_g,C_h)=\sum a_{ij}$,其中 $i\in C_g$,$j\in C_h$。非规范化谱聚类的目标是将图 G 划分为 k 个簇,并最小化非规范化谱聚类目标函数,非规范化谱聚类目标函数[35](Ratio-Cut)的定义如下:

$$RatioCut(C_1,C_2,\cdots,C_k)=\min\left(\sum_{j=1}^{k}\frac{A(C_j,\overline{C_j})}{\sqrt{|C_j|}}\right) \tag{1}$$

文献[36]证明了图拉普拉斯矩阵的特征向量与图 G 的分割有关(在松弛意义上)。因此,Ratio-Cut 目标函数可以写成:

$$\begin{aligned}&Ratio(C_1,C_2,\cdots,C_k)=\operatorname{argmin}\ \boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\\&\text{s.t. }\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\mathbf{I}_k\end{aligned} \tag{2}$$

其中,$\boldsymbol{H}\in\mathbb{R}^{n\times k}$是一个松弛的簇指示矩阵,如果节点 i 在子图 C_j 中,那么 $\boldsymbol{H}_{ij}=1/\sqrt{|C_j|}$,否则 $\boldsymbol{H}_{ij}=0$。$\boldsymbol{L}$ 是图 G 的拉普拉斯矩阵,$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$。矩阵 $\boldsymbol{D}=diag(d_{11},\cdots d_{m})$被称为度矩阵,其中 $d_{ii}=\sum_{j=1}^{n}a_{ij}$。$\boldsymbol{A}=(a_{ij})_{n\times n}$是图 G 的相似矩阵。

在公平聚类任务中,通常假设在数据集上存在两组属性,一组是聚类兴趣属性(Clustering interest Attribute,CA),另一组是敏感属性(Sensitive Attributes,SA)。其中 CA 属性集是与聚类任务相关的属性,而 SA 属性集则是希望在算法输出中保持公平的属性集。实现公平聚类的一种自然方式是确保每个簇中保护组的分布近似于其在数据集中的分布,这与公平监督学习中被广泛使用的对等统计概念[6]有关。我们将具有公平性约束的非规范化谱聚类问题定义为定义 1。

定义 1(非规范化公平谱聚类问题) 给定 l 个组 $P_1,P_2,\cdots,P_l$,作为敏感属性的值,$P_1\cup P_2\cup\cdots\cup P_l=P$ 且 $P_i\cap P_j=\emptyset,i,j\in[l],i\neq j$。公平谱聚类问题可以描述为寻找图 G 的一个分割,使得最小化非规范化的谱聚类目标函数满足如下公平约束:

$$\frac{|C_j\cap P_s|}{|C_j|}=\frac{|P_s|}{|P|},j\in[k],s\in[l] \tag{3}$$

4 公平谱聚类方法

本节首先介绍如何将公平约束整合到非规范化的谱聚类;接着介绍本文所提出的两种公平谱聚类方法,为了方便理解,先介绍单个敏感属性的公平谱聚类方法,然后从单个敏感属性公平谱方法过渡到多个敏感属性的公平谱方法,最后分析所提算法的复杂度。

4.1 将公平约束整合到谱聚类

在公平聚类结果中,保护组在每个簇中的分布被期望近似于其在数据集中的分布。为了放宽公平约束限制,我们在定义 1 的公平约束中引入了公平上下界[14]的概念。具体来说,我们引入了一个超参数 α,用于控制公平约束的上下界,保护组如果在每个簇中的分布处于设定的上下界内,则认为聚类是公平的。在引入超参数 α 后,等式(3)中的公平约束可以用式(4)表示:

$$\frac{|P_s|}{|P|}-\frac{|C_j\cap P_s|}{|C_j|}\leqslant|\alpha|,j\in[k],l\in[l] \tag{4}$$

令 $\boldsymbol{R}=(r_{is})_{n\times l}$表示保护组的指示矩阵,其中元素 r_{is} 的定义如下:

$$r_{is}=\begin{cases}1, & i\in P_s\\0, & i\notin P_s\end{cases} \tag{5}$$

对于任意节点 i 属于一个具有 l 值的敏感属性时,其指示矩阵的行向量满足 $r_{i1}+r_{i2}+\cdots+r_{il}=1$。

定理 1 用指示矩阵 $\boldsymbol{R}$ 和 $\boldsymbol{H}$ 对约束条件进行编码,则式(4)可被转化为如下形式:

$$\sum_{i=1}^{n}\left[\left(\frac{|P_s|}{|P|}-r_{is}\right)h_{ij}\right]<|\alpha|,s\in[l] \tag{6}$$

证明:式(4)转化为式(6)的证明如下:

$$\frac{|P_s|}{|P|}-\frac{|C_j\cap P_s|}{|C_j|}=\frac{|P_s||C_j|}{|P|\sqrt{|C_j|}}-\frac{|C_j\cap P_s|}{\sqrt{|C_j|}} \tag{7}$$

由指示向量 $\boldsymbol{r}$ 和指示向量 $\boldsymbol{h}$ 的定义可知:

$$r_{is}h_{ij}=\begin{cases}\dfrac{1}{\sqrt{|C_j|}}, & i\in P_s\cap C_j\\0, & i\notin P_s\cap C_j\end{cases}$$

因此,有:

$$\frac{|P_s||C_j|}{|P|\sqrt{|C_j|}}-\frac{|C_j\cap P_s|}{\sqrt{|C_j|}}=\sum_{i=1}^{n}h_{ij}\frac{|P_s|}{|P|}-\sum_{i=1}^{n}h_{ij}r_{is}$$

令 $\boldsymbol{F}_{n\times l}$表示一个以 $\mathbf{1}_n\cdot|P_s|/|P|-r_{is}$ 为列的约束矩阵,则式(4)可以进一步地用式(7)来表示:

$$\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\leqslant|\alpha| \tag{7}$$

将式(7)中矩阵形式的公平概念视为非规范化的谱聚类约束条件,并将其整合到非规范化谱聚类的目标函数(见式(2)),有:

$$\begin{aligned}&RatioCut(C_1,C_2,\cdots,C_k)=\operatorname{argmin}\ \boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\\&\text{s.t. }\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\mathbf{I}_k,\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\leqslant|\alpha|_{l\times k}\end{aligned} \tag{8}$$

因此,公平谱聚类问题转化成了式(8)中的约束优化问题。

4.2 非规范化公平谱聚类

在解决一个约束优化问题时,通常采用的是卡罗需-库恩-塔克(Karush-Kuhn-Tucker,KKT)定理[37],KKT 定理规定了约束优化问题取得最优解的必要条件。我们可以得到一组满足所有 KKT 条件的候选解,其也被称为可行解。在可行集较小的情况下,利用蛮力法在可行解中寻找最优解。

首先,在式(8)目标函数中引入拉格朗日乘子,构建拉格朗日函数如下:

$$\begin{aligned}\mathbb{L}(\boldsymbol{H},\lambda_1,\lambda_2,\mu)=&\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}+\lambda_1(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}-\alpha)-\lambda_2(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}+\alpha)+\\&\mu(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}-\mathbf{I}_k)\end{aligned} \tag{9}$$

根据 KKT 定理可知,式(8)中的任意可行解都需要满足如下条件:

$$\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}+\lambda_1\ \boldsymbol{F}^{\mathrm{T}}-\lambda_2\ \boldsymbol{F}^{\mathrm{T}}+\mu\boldsymbol{H}^{\mathrm{T}}=0 \tag{10}$$

$$\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}-\alpha\leqslant0,\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}+\alpha\geqslant0,\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\mathbf{I}_k \tag{11}$$

$$\lambda_1\geqslant0,\lambda_2\geqslant0 \tag{12}$$

$$\lambda_1(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}-\alpha)=0,\lambda_2(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}-\alpha)=0 \tag{13}$$

其中,式(10)是式(9)对指示矩阵 $\boldsymbol{H}$ 求导后的结果。式(10)被称为平稳性条件,式(11)是初步可行性条件,式(12)被称为双重可行性条件,式(13)是松弛互补条件。

观察式(13)中的松弛互补条件可以发现,它可以分成 4 种情况。我们分别在这 4 种情况下对式(10)、式(13)进行求解。

情况 1 $\lambda_1=0,\lambda_2=0$。此时 KKT 条件变成:

$$\begin{aligned}&\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}+\mu\boldsymbol{H}^{\mathrm{T}}=0\\&-\alpha\leqslant\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\leqslant\alpha\\&\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\mathbf{I}_k\end{aligned} \tag{14}$$

从式(14)中容易知道,在 $\lambda_1=0,\lambda_2=0$ 的情况下产生的

可行解一定是拉普拉斯矩阵 $\boldsymbol{L}$ 的特征向量。因此,我们只需要求出 $\boldsymbol{L}$ 的特征向量并去除不满足初步可行性条件 $-\alpha \leqslant \boldsymbol{F}^{\mathrm{T}}\boldsymbol{H} \leqslant \alpha$ 的特征向量,即可得到可行解。需要注意的是,通过改变超参数 α 的值可以控制满足条件的可行解数量。

情况 2　$\lambda_1 \neq 0, \lambda_2 \neq 0$。此时 KKT 条件变成:

$$\begin{aligned}&\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}+\lambda_1\ \boldsymbol{F}^{\mathrm{T}}-\lambda_2\ \boldsymbol{F}^{\mathrm{T}}+\mu\boldsymbol{H}^{\mathrm{T}}=0\\&\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}=0, \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\mathrm{I}_k\\&\lambda_1>0, \lambda_2>0\end{aligned} \tag{15}$$

然而,在这种情况下式(15)并不容易求解。由式(15)的初步可行性条件可以看出,情况 2 的可行解一定是式(15)的解。

$$\begin{aligned}&RatioCut(C_1, C_2, \cdots, C_k)=\operatorname{argmin}\ \boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\\&\text{s. t. }\ \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\mathbf{I}_k, \boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}=\mathbf{0}_{l\times k}\end{aligned} \tag{16}$$

本文采用文献[38]中讨论的一种方法来求解式(16)。令 $\boldsymbol{Z}$ 是一个矩阵,它的列向量构成了 $\boldsymbol{F}^{\mathrm{T}}$ 的零空间的正交基。令 $\boldsymbol{H}=\boldsymbol{Z}\boldsymbol{Y}$,那么式(16)可转化为:

$$\begin{aligned}&RatioCut(C_1, C_2, \cdots, C_k)=\operatorname{argmin}\ (\boldsymbol{Y}^{\mathrm{T}}\ \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{Z}\boldsymbol{Y})\\&\text{s. t. }\ \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}=\mathbf{I}_k\end{aligned} \tag{17}$$

因此,求等式(15)的可行解可以看作是求解等式(17)。与等式(2)类似,等式(17)的解构成了矩阵 $\boldsymbol{Y}$。其中,$\boldsymbol{Y}$ 的列向量是矩阵 $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{Z}$ 的 k 个最小特征值对应的特征向量。然后,我们令 $\boldsymbol{H}=\boldsymbol{Z}\boldsymbol{Y}$ 得到等式(15)的可行解。

情况 3　$\lambda_1 = 0, \lambda_2 \neq 0$。此时 KKT 条件变成:

$$\begin{aligned}&\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}-\lambda_2\ \boldsymbol{F}^{\mathrm{T}}+\mu\boldsymbol{H}^{\mathrm{T}}=0\\&\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}-\alpha=0, \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\mathbf{I}_k\\&\lambda_2>0\end{aligned} \tag{18}$$

显然,这种情况也可以通过情况 2 中的方法求解。

情况 4　$\lambda_1 \neq 0, \lambda_2 = 0$。与情况 3 类似。

通过求解以上 4 种情况下的可行解,我们得到等式(8)的一个可行解集。然后,寻找可行解集中的最优解,从而实现公平谱聚类。具体的算法如算法 1 所示。

算法 1　非规范化公平谱聚类

输入:邻接矩阵 **A**,聚类的簇数 k,松弛参数 α,敏感属性的约束条件 r

输出:最优聚类指示向量 $\mathbf{H}^*$

1. 利用度矩阵 **D** 计算拉普拉斯矩阵 $\mathbf{L}=\mathbf{D}-\mathbf{A}$
2. 构建以 $1_n \cdot |P_s|/|P| - r_{is}$ 位列向量的约束矩阵 **F**
3. 计算拉普拉斯矩阵 **L** 的特征值并保留满足约束条件 $-\alpha \leqslant \mathbf{F}^{\mathrm{T}}\ \mathbf{D}^{-1/2}\mathrm{v} \leqslant \alpha$ 的特征向量作为可行解
4. 计算矩阵 **Z**,它的列向量构成 $\mathbf{F}^{\mathrm{T}}$ 零空间的正交基
5. 计算矩阵 **Y**,它的列向量由矩阵 $\mathbf{Z}^{\mathrm{T}}\mathbf{L}\mathbf{Z}$ 的 k 个最小特征值对应的特征向量构成
6. 计算特征向量 $\mathbf{H}=\mathbf{Z}\mathbf{Y}$,并将其加入可行解集
7. 返回可行解中的最优解 $\mathbf{H}^*$

4.3　多个敏感属性公平谱聚类

4.2 节介绍了非规范化的公平谱聚类方法如何在单个敏感属性的条件下工作。现在,我们将它扩展到多个敏感属性的情况。

与单个敏感属性不同的是,多个敏感属性公平约束条件增多,从而改变等式(8)中的约束条件的个数。假设存在 φ 个敏感属性,那么等式(8)中非规范化公平谱聚类的目标函数可写为如下形式:

$$\begin{aligned}&RatioCut(C_1, C_2, \cdots, C_k)=\operatorname{argmin}\ \boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\\&\text{s. t. }\ \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\mathbf{I}_k\\&\boldsymbol{F}_1^{\mathrm{T}}\boldsymbol{H} \leqslant |\alpha|_{\varphi l\times k}, \boldsymbol{F}_2^{\mathrm{T}}\boldsymbol{H} \leqslant |\alpha|_{\varphi l\times k}, \cdots, \boldsymbol{F}_{\varphi}^{\mathrm{T}}\ \boldsymbol{H} \leqslant |\alpha|_{\varphi l\times k}\end{aligned} \tag{19}$$

同样地,对于一个约束优化问题,我们可以在等式(19)的目标函数中引入拉格朗日乘子,并使用 KKT 定理求其可行解集,从中找到最优解。但是,随着约束条件的增多,使用 KKT 定理求可行解会变得非常复杂。因此,我们增加公平约束的严格性,将非规范化公平谱聚类的目标函数改写为:

$$\begin{aligned}&RatioCut(C_1, C_2, \cdots, C_k)=\operatorname{argmin}\ \boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\\&\text{s. t. }\ \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\mathbf{I}_k\\&\boldsymbol{F}_1^{\mathrm{T}}\boldsymbol{H}=\mathbf{0}_{\varphi l\times k}, \boldsymbol{F}_2^{\mathrm{T}}\boldsymbol{H}=\mathbf{0}_{\varphi l\times k}, \cdots, \boldsymbol{F}_{\varphi}^{\mathrm{T}}\ \boldsymbol{H}=\mathbf{0}_{\varphi l\times k}\end{aligned} \tag{20}$$

为了简化等式(20)的表达式,我们令 $\boldsymbol{F}_m$ 表示 φ 个敏感属性的约束矩阵,$\boldsymbol{F}_m$ 的定义如下:

$$\boldsymbol{F}_m=\begin{bmatrix}\boldsymbol{F}_1\\ \boldsymbol{F}_2\\ \vdots\\ \boldsymbol{F}_{\varphi}\end{bmatrix} \tag{21}$$

然后,令:

$$\boldsymbol{F}_m^{\cdot\,\mathrm{T}}=\begin{bmatrix}\boldsymbol{F}_1^{\mathrm{T}}\\ \boldsymbol{F}_2^{\mathrm{T}}\\ \vdots\\ \boldsymbol{F}_{\varphi}^{\mathrm{T}}\end{bmatrix} \tag{22}$$

表示约束矩阵的转置,则等式(20)中的目标函数可写为:

$$\begin{aligned}&RatioCut(C_1, C_2, \cdots, C_k)=\operatorname{argmin}\ \boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\\&\text{s. t. }\ \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\mathbf{I}_k, \boldsymbol{F}_m^{\cdot\,\mathrm{T}}\boldsymbol{H}=\mathbf{0}_{\varphi l\times k}\end{aligned} \tag{23}$$

由分块矩阵的性质可知,对于矩阵 $\boldsymbol{A}$ 和矩阵 $\boldsymbol{B}$,它们的零空间 $N(\boldsymbol{A})$ 和 $N(\boldsymbol{B})$ 的交集等于矩阵 $\boldsymbol{C}$ 的零空间 $N(\boldsymbol{C})$,其中 $\boldsymbol{C}$ 是一个分块矩阵,并且 $\boldsymbol{C}^{\mathrm{T}}=[\boldsymbol{A}\ \boldsymbol{B}]$。因此,对于等式(23),我们可以使用 2.3 节中情况 2 的方法来求解。即令 $\boldsymbol{Z}_m$ 是一个矩阵,它的列向量构成了 $\boldsymbol{F}_m^{\cdot\,\mathrm{T}}$ 的零空间的正交基。我们令 $\boldsymbol{H}=\boldsymbol{Z}_m\ \boldsymbol{Y}_m$,则等式(23)可写成:

$$\begin{aligned}&RatioCut(C_1, C_2, \cdots, C_k)=\operatorname{argmin}\ (\boldsymbol{Y}_m^{\mathrm{T}}\ \boldsymbol{Z}_m^{\mathrm{T}}\boldsymbol{L}\boldsymbol{Z}_m\ \boldsymbol{Y}_m)\\&\text{s. t. }\ \boldsymbol{Y}_m^{\mathrm{T}}\ \boldsymbol{Y}_m=\mathbf{I}_k\end{aligned} \tag{24}$$

然后,由 $\boldsymbol{H}=\boldsymbol{Z}_m\ \boldsymbol{Y}_m$ 得到目标函数的解。通过求解等式(23),从而实现了多敏感属性条件下的公平谱聚类方法。具体的算法如算法 2 所示。

算法 2　多敏感属性公平谱聚类

输入:邻接矩阵 **A**,聚类的簇数 k,敏感属性的约束条件 $\mathrm{r_m}=\{\mathrm{r_1}, \mathrm{r_2}, \cdots, \mathrm{r_m}\}$

输出:最优聚类指示向量 $\mathbf{H}_{\mathrm{m}}^*$

1. 利用度矩阵 **D** 计算拉普拉斯矩阵 $\mathbf{L}=\mathbf{D}-\mathbf{A}$
2. 构建以 $1_n \cdot |P_s|/|P| - r_{is}$ 为列向量的约束矩阵 $\mathbf{F}_i$,$i \in [\varphi]$ 和矩阵约束 $\mathbf{F}_{\mathrm{m}}$
3. 计算矩阵 $\mathbf{Z}_{\mathrm{m}}$,它的列向量构成零空间的正交基
4. 计算矩阵 $\mathbf{Y}_{\mathrm{m}}$,它的列向量由矩阵 $\mathbf{Z}_{\mathrm{m}}^{\mathrm{T}}\mathbf{L}\ \mathbf{Z}_{\mathrm{m}}$ 的 k 个最小特征值对应的特征向量构成
5. 返回 $\mathbf{H}_{\mathrm{m}}^* = \mathbf{Z}\mathbf{Y}$

4.4　复杂度分析

在标准的谱聚类算法中,算法的复杂度主要是由计算

特征向量的复杂度决定。一般来说，对于任意的聚类簇数 k，谱聚类的算法复杂度均为 $O(n^3)$。除了一般谱聚类需要进行的计算外，在算法 1 中需要移除不满足约束条件 $-\alpha \leqslant \boldsymbol{F}^{\mathrm{T}}\boldsymbol{H} \leqslant \alpha$ 的特征向量，这一步的复杂度为 $O(ln^2)$。此外，算法 1 还需要计算 $\boldsymbol{F}^T$ 零空间的正交基。利用奇异值分解(Singular Value Decomposition，SVD)可以计算 $\boldsymbol{F}^{\mathrm{T}}$ 的零空间的正交基，SVD 可以在 $O(n^3)$内完成。因此，算法 1 的复杂度为 $O(n^3)$。在算法 2 中，则只需要使用奇异值分解计算 $\boldsymbol{F}_m^{\mathrm{T}}$ 零空间的正交基，所以算法 2 的复杂度为 $O(n^3)$。

5 实验与结果

本节展示了本文提出的两种公平聚类方法，即非规范化公平谱聚类方法(UFSC)和多敏感属性公平谱聚类方法(MFSC)在真实数据集上的实验内容。首先概述了实验中使用的数据集和对比算法，然后描述了实验所用的评测指标，最后报告了实验的结果。

5.1 数据集和实验设置

本文在实验中使用了 5 个公平聚类领域内常用的数据集。1)Drug[39]。Drug 数据集是根据哈特福德吸毒者社交关系编码的一个网络，它包含 286 个节点。2)Adult[40]。Adult 数据集也被称为 census，它包含了 1994 年美国人口普查的信息，我们将其采样为 600 条信息。3)Obesity[41]。Obesity 数据集是根据墨西哥、秘鲁和哥伦比亚等国家的个人饮食习惯和身体状况估计其肥胖水平的数据，它包含 2 111 条数据。4)Bank[42]。Bank 数据集中包含与葡萄牙银行机构的电话营销活动有关的记录，我们将其采样为 1 000 条数据。5)Census1990[43]。Census1990 是 1990 年的美国人口普查记录，它包含 2 458 285 条记录，我们在此数据集上进行运行时间分析。对于每一个数据集，我们都选择了数字属性作为聚类兴趣属性(CA)。此外，我们还为每个数据集设置了敏感属性(SA)，并根据其取值创建了保护组。有关数据集设置的更多信息如表 1 所列。

表 1 聚类兴趣属性和敏感属性

Table 1 Clustering interest attributes and sensitive attributes

Datasets	Clustering Attribute	Sensitive Attribute	Group
Drug	Acquaintanceship matrix	Gender	male, female
		Ethnicity	African American, white or other, Puerto Rican/Latino
Adult	age, education-num, num-medications, num-outpatient, num-emergency, num-inpatient	sex	female, male
		income	>50k, <=50k
Obesity	Age, Height, Weight, FAVC, FCVC, NCP, CAEC, SMOKE, CH2OSCC, FAF, TUE, CALC, MTRANS	Gender	male, female
		family_history	yes, no
Bank	age, balance, duration	marital	married, single, divorced
		default	yes, no
Census1990	dAncstry1, dAncstry2, iAvail, iCitizen, iClass, dDepart, iFertil, iDisabl1, iDisabl2, iEnglish, iFeb55, dHispanic, dHour89	iSex	male, female
		dAge	8 groups

为了观察本文算法的性能，我们使用两种对比算法与本文算法进行比较。对比算法包括：1) 非正则化谱聚类方法[35](Unnormalized Spectral Clustering，SC)；2) 带有公平约束的谱聚类方法[18](Spectral Clustering with Fairness Constraints，SCFC)。

在实验中，簇的数目被设置为 5～20，以观察在不同聚类簇数下本文方法的性能。此外，所有的实验均运行 20 次，并取其平均结果以避免随机性对实验结果的影响。对于超参数 α，我们建议根据保护组在数据集中所占的比例进行设置。在 Bank 数据集的 marital，default 属性以及 Drug 数据集的 Ethnicity 属性的实验中，α 被设置为 0.1，在其他实验中 α 被设置为 0.04。

5.2 评测指标

我们使用保护组分布向量的平均欧几里得距离[21](Average Euclidean Distance，AED)来测量簇中的不公平性。令 SA 是一个敏感属性，它有 l 个取值(l 个保护组)。这些保护组在数据集 P 中的分布产生了一个长度为 l 的分布向量 $\boldsymbol{P}_{\mathrm{SA}}$。同样，这些保护组在每个簇中的分布产生长度为 l 的分布向量 $\boldsymbol{C}_{\mathrm{SA}}$。通过计算数据集中的 $\boldsymbol{P}_{\mathrm{SA}}$ 和聚类结果中的 $\boldsymbol{C}_{\mathrm{SA}}$ 之间的欧氏距离，可以得到分布向量之间的欧几里得距离 ED。AED 是每个簇中分布向量的平均欧几里得距离，即 $AED=ED/k$。其中，k 是聚类中簇的数目。一个聚类中 AED 指标的值越低，则表示聚类越公平。

5.3 非规范化公平谱聚类

首先对 UFSC 的性能进行了评估，AED 指标被用来测量聚类的公平性(AED 的定义见 3.2 节)，实验结果如图 1 所示。图 1 中的 Y 轴表示公平性指标 AED 的值，X 轴表示簇的数量，不同算法的结果在图中以不同颜色标识。从图中可以看出，UFSC 的 AED 指标在所有的实验中均低于基线算法 SC 的 AED 指标，而对比算法 SCFC 则并不能保证得到一个比 SC 更加公平的聚类结果。这说明 UFSC 提供了比 SCFC 更强的公平保证。表 2 列出了 UFSC 和 SCFC 在不同数据集中聚类结果的 AED 指标，最优的结果均被加粗显示。其中，每个数据集中的 AED 指标的值均取的是对应数据集所有实验的均值。从表 2 可以看出，与 SCFC 算法相比，UFSC 在 Bank 数据集上的表现最好，提升了 47.99%的公平性，在 Drug 数据集上 UFSC 提升了 14.61% 的公平性，在 Obesity 数据集上提升了 14.61%的公平性。UFSC 在 Adult 数据集上表现较差，但仍提升了 11.58%的公平性。总的来说，UFSC 聚类结果中的公平性比 SCFC 聚类结果的公平性提升了 22.11%。图 1 和表 2 所示的实验结果证明了本文提出的公平聚类公式在单个敏感属性条件下对提升聚类公平程度的有效性。

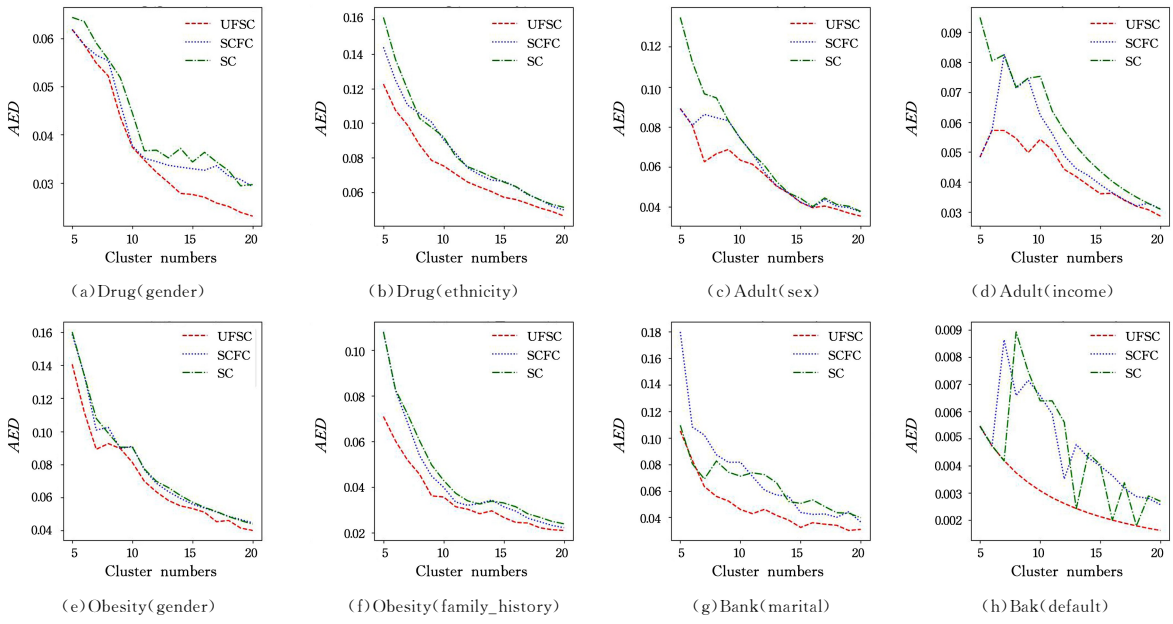

图 1　UFSC 与 SCFC,SC 的公平性比较(电子版为彩图)

Fig. 1　Fairness comparison between UFSC and SCFC,SC

表 2　UFSC 与 SCFC 的平均公平性

Table 2　Average fairness of UFSC and SCFC

methods	datasets			
	Drug	Adult	Obesity	Bank
UFSC	0.0534	0.0492	0.0527	0.0256
SCFC	0.0612	0.0549	0.0604	0.0378

5.4　多敏感属性公平谱聚类

我们同时使用数据集中设定的多个敏感属性(而不是像 5.3 节单独使用)来评估 MFSC 方法的性能,图 2 给出了实验结果。

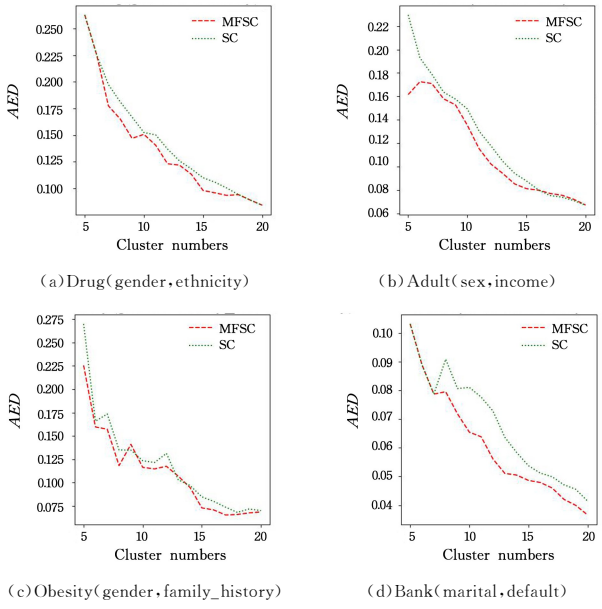

图 2　MFSC 与 SC 的公平性比较(电子版为彩图)

Fig. 2　Fairness comparison between MFSC and SC

与图 1 类似,图 2 中,Y 轴用于表示聚类结果中 AED 指标的值,X 轴用于表示簇的个数,不同算法的结果在图中以不同颜色标识。由于 SCFC 算法不支持在多敏感属性条件下工作,因此采用 SC 作为对比方法。从图 2 中可以看出,MFSC 的聚类结果比 SC 的聚类结果有更低的 AED 值。这说明了在多个敏感属性的条件下,MFSC 能够寻找到比 SC 更加公平的聚类。由于 MFSC 是 SCFC 算法在多敏感属性条件下的一种推广方法,因此 MFSC 不能提供很好的公平保证。虽然 MFSC 是第一个实现多敏感属性的公平谱聚类方法,但提升 MFSC 的公平性表现确实是一个值得改进的方面。

5.5　运行时间分析

在图 3 所示的实验中,我们分析了在不同数量的数据点下本文提出的两种方法(UFSC 和 MFSC)的运行时间。

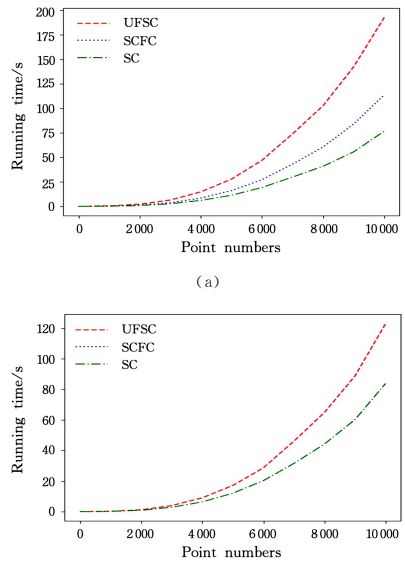

图 3　运行时间

Fig. 3　Running time

图 3 中,Y 轴表示算法生成聚类结果所运行的时间,X 轴表示实验中使用的数据量,不同算法的运行时间用不同颜色标识。从图 3(a)可以看出,UFSC 生成聚类结果所消耗的

时间比 SCFC 和 SC 都长。UFSC 的运行时间约是 SC 算法的 2.47 倍，是 SCFC 算法的 1.76 倍。从图 3(b)可以看出，MFSC 的运行时间同样长于 SC 算法，约是 SC 算法的 1.42 倍。尽管 UFSC 和 MFSC 的运行时间稍长于其他对比算法，但考虑到 UFSC 和 MFSC 提供了更高的聚类公平性，我们认为这是可接受的。

结束语 本文研究了谱聚类算法的公平性问题，并提出了一种非规范化的公平谱聚类方法和一种适用于多个敏感属性的公平谱聚类方法。本文将公平性概念整合为约束矩阵，从而将公平性问题转化为约束谱聚类问题。通过解决公平约束下的谱聚类问题，实现了公平聚类。在多个真实数据集上的实验证明，本文方法相比其他对比算法提供了更加公平的聚类结果。

由于 MFSC 不能提供很好的公平保证，我们未来工作的一个重要方向是提升 MSFC 的公平性能，使得 MFSC 的聚类结果更加公平。此外，本文仅给出了非规范化的公平谱聚类方法，开发规范化谱聚类的公平版本也是一个重要的研究方向。

参 考 文 献

[1] USHIODA A. Hierarchical clustering of words and application to NLP tasks[C]// Fourth Workshop on Very Large Corpora. 1996.

[2] WU Z, LEAHY R. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1993, 15(11): 1101-1113.

[3] HUTTENHOWER C, FLAMHOLZ A I, LANDIS J N, et al. NearestNeighbor Networks: clustering expression data based on gene neighborhoods[J]. BMC Bioinformatics, 2007, 8(1): 1-13.

[4] CHUANG K H, CHIU M J, LIN C C, et al. Model-free functional MRI analysis using Kohonen clustering neural network and fuzzy C-means[J]. IEEE Transactions on Medical Imaging, 1999, 18(12): 1117-1128.

[5] IYER N S, KANDEL A, SCHNEIDER M. Feature-based fuzzy classification for interpretation of mammograms[J]. Fuzzy Sets and Systems, 2000, 114(2): 271-280.

[6] DWORK C, HARDT M, PITASSI T, et al. Fairness through awareness[C]// Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. New York: ACM, 2012: 214-226.

[7] KLEINBERG J, LAKKARAJU H, LESKOVEC J, et al. Human decisions and machine predictions[J]. The Quarterly Journal of Economics, 2018, 133(1): 237-293.

[8] ROMEI A, RUGGIERI S. A multidisciplinary survey on discrimination analysis[J]. The Knowledge Engineering Review, 2014, 29(5): 582-638.

[9] FELDMAN M, FRIEDLER S A, MOELLER J, et al. Certifying and removing disparate impact[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2015: 259-268.

[10] HARDT M, PRICE E, SREBRO N. Equality of opportunity in supervised learning[C]// Advances in Neural Information Processing Systems. 2016: 3315-3323.

[11] NARAYANAN A. Translation tutorial: 21 fairness definitions and their politics[C]// Proceedings of the Conference on Fairness Accountability Transp. New York, USA, 2018.

[12] ZAFAR M B, VALERA I, GOMEZ RODRIGUEZ M, et al. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment[C]// Proceedings of the 26th International Conference on World Wide Web. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017: 1171-1180.

[13] HASHIMOTO T, SRIVASTAVA M, NAMKOONG H, et al. Fairness without demographics in repeated loss minimization [C]// International Conference on Machine Learning. PMLR, 2018: 1929-1938.

[14] CHIERICHETTI F, KUMAR R, LATTANZI S, et al. Fair clustering through fairlets[C]// Advances in Neural Information Processing Systems. 2017: 5029-5037.

[15] RÖSNER C, SCHMIDT M. Privacy Preserving Clustering with Constraints[C]// 45th International Colloquium on Automata, Languages, and Programming (ICALP 2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[16] BERCEA I O, GROß M, KHULLER S, et al. On the Cost of Essentially Fair Clusterings[C]// Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[17] BACKURS A, INDYK P, ONAK K, et al. Scalable fair clustering[C]// International Conference on Machine Learning. PMLR, 2019: 405-413.

[18] KLEINDESSNER M, SAMADI S, AWASTHI P, et al. Guarantees for spectral clustering with fairness constraints[C]// International Conference on Machine Learning. PMLR, 2019: 3458-3467.

[19] BERA S K, CHAKRABARTY D, FLORES N J, et al. Fair algorithms for clustering[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 4954-4965.

[20] AHMADIAN S, EPASTO A, KUMAR R, et al. Clustering without over-representation[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2019: 267-275.

[21] ABRAHAM S S, PADMANABHAN D, SUNDARAM S S. Fairness in Clustering with Multiple Sensitive Attributes[C]// EDBT/ICDT 2020 Joint Conference. 2020: 287-298.

[22] SCHMIDT M, SCHWIEGELSHOHN C, SOHLER C. Fair coresets and streaming algorithms for fair k-means[C]// International Workshop on Approximation and Online Algorithms. Cham: Springer, 2019: 232-251.

[23] AHMADIAN S, EPASTO A, KUMAR R, et al. Fair correlation clustering[C]// International Conference on Artificial Intelligence and Statistics. PMLR, 2020: 4195-4205.

[24] KLEINDESSNER M, AWASTHI P, MORGENSTERN J. Fair k-center clustering for data summarization[C]// International

Conference on Machine Learning. PMLR,2019:3448-3457.
[25] ZIKO I M,GRANGER E,YUAN J,et al. Clustering with fairness constraints: A flexible and scalable approach[J]. arXiv:1906.08207,2019.
[26] CHEN X,FAIN B,LYU L,et al. Proportionally fair clustering[C]// International Conference on Machine Learning. PMLR,2019:1032-1041.
[27] JUNG C,KANNAN S,LUTZ N. Service in Your Neighborhood: Fairness in Center Location[C]// 1st Symposium on Foundations of Responsible Computing(FORC 2020). Schloss Dagstuhl-Leibniz-Zentrum für Informatik,2020.
[28] DAVIDSON I,RAVI S S. Making existing clusterings fairer: Algorithms,complexity results and insights[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 3733-3740.
[29] GHADIRI M,SAMADI S,VEMPALA S. Socially fair k-means clustering[C]// Proceedings of the 2021 ACM Conference on Fairness,Accountability,and Transparency. 2021:438-448.
[30] ABBASI M,BHASKARA A,VENKATASUBRAMANIAN S. Fair clustering via equitable group representations[C]// Proceedings of the 2021 ACM Conference on Fairness,Accountability,and Transparency. 2021:504-514.
[31] MAKARYCHEV Y,VAKILIAN A. Approximation Algorithms for Socially Fair Clustering[J]. arXiv:2103.02512,2021.
[32] CHHABRA A,MOHAPATRA P. Fair algorithms for hierarchical agglomerative clustering[J]. arXiv:2005.03197,2020.
[33] AHMADIAN S,EPASTO A,KNITTEL M,et al. Fair Hierarchical Clustering[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020: 21050-21060.
[34] QUY T L,ROY A,FRIEGE G,et al. Fair-Capacitated Clustering[J]. arXiv:2104.12116,2021.
[35] VON LUXBURG U. A tutorial on spectral clustering[J]. Statistics and Computing,2007,17(4):395-416.
[36] YU S X,SHI J. Segmentation given partial grouping constraints[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2004,26(2):173-183.
[37] BERTSEKAS D P. Nonlinear programming[J]. Journal of the Operational Research Society,1997,48(3):334-334.
[38] SHI J,MALIK J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(8):888-905.
[39] WEEKS M R,CLAIR S,BORGATTI S P,et al. Social networks of drug users in high-risk sites: Finding the connections[J]. AIDS and Behavior,2002,6(2):193-206.
[40] ZHOU Z H,CHEN Z Q. Hybrid decision tree[J]. Knowledge-based systems,2002,15(8):515-528.
[41] PALECHOR F M,DE LA HOZ MANOTAS A. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia,Peru and Mexico[J]. Data in Brief,2019,25:104344.
[42] MORO S,CORTEZ P,RITA P. A data-driven approach to predict the success of bank telemarketing[J]. Decision Support Systems,2014,62:22-31.
[43] MEEK C,THIESSON B,HECKERMAN D. The learning-curve sampling method applied to model-based clustering[J]. Journal of Machine Learning Research,2002,2(2):397-418.

XU Xia, born in 1996, postgraduate, is a member of China Computer Federation. His main research interests include fair clustering, spectral clustering and fair machine learning.

ZHANG Hui, born in 1972, Ph.D, professor, is a member of China Computer Federation. His main research interests include big data and machine learning.

(责任编辑:喻藜)