文章编号:1000-5862(2021)03-0292-07

基于几何特征的学生评教数据离群点检测算法

唐宇坤,邓 松*,许梦雅,郭 馨
(江西财经大学软件与物联网工程学院,江西 南昌 330013)

摘要:针对学生评教数据中的离群点问题,根据消极评教数据产生的方式及特点,提出了一种基于几何特征的学生评教数据离群点检测算法. 该算法通过分析样本的几何特征,计算样本的离群程度,完成离群点检测,共分为 3 步进行:(i)依据教学质量评价数据,在几何特征空间中建立样本的点映射;(ii)从形状相似度、距离相似度 2 个方面构建判别空间,对几何特征空间中的样本点进行分析运算,得到样本点在判别空间中的点映射;(iii)以基于半监督近邻的方法对判别空间中的样本进行检测. 实验结果表明:该算法检测精度较高,在高校教师教学效果中有较好的应用价值.

关键词:学生评教;几何特征;离群点;支持向量机
中图分类号:TP 311 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2021.03.11

0 引言

学生评教是学生按照设定的评价项目,对所学课程的教师授课等教学状况进行评分的活动. 通过统计学生在每个评价项目上的评分,计算出教师的最终得分,并将此分数作为教师教学水平的体现. 学生消极评教产生的离群数据,会使教学质量评价工作产生误用性风险[1],即可能出现教师的最终得分无法体现教师真实教学水平的情况,因此需要设计算法减少离群数据带来的影响.

为便于分析,根据学生评教数据特点,结合以下 2 个实例进行分析.

例 1 在某教师的学生评教数据中,学生 A 填写的各项评教项目分数与班级总体的评分趋势相违. 大部分学生评价高分的项目,学生 A 评出低分;大部分学生评价低分的项目,学生 A 评出高分. 同时,学生 A 的评价项目总分与班级平均评分相近.

例 2 在某教师的学生评教数据中,学生 B 填写的各项评教项目分数与班级总体的项目评分水平相比偏高. 如大部分学生评出 6 分的评价项目,学生 B 评出 8 分;大部分学生评出 7 分的评价项目,学生 B 评出 9 分.

从以上 2 个例子可以看出,在学生评教数据中的虚假数据与目前常用的风险评价模型存在一定的区别.

在例 1 中,学生 A 评出的各项目分数,尽管与总体的评分趋势相违,但各评价项目得分偏差相抵,计算得出的最终分数仍与总体的评分水平相符,具有一定的隐蔽性. 在教学评价工作中,需根据总体学生的评分状况,设置各评价项目权重. 在例 1 中的学生 A 虽然评出的总分与总体评分水平相符,但是目前高校往往要根据总体评分情况对各评分项目进行权重计算[2],在项目权重重新计算后,此时学生 A 评出的总分会与总体评分水平出现较大偏差. 而且学生 A 对各项评价项目的评分也会对评价项目权重设置算法带来一定的影响.

在例 2 中,学生 B 在有意或无意中评出了偏高的评分. 尽管存在学生 B 没有进行消极评教的可能性,但是学生 B 的评分无法反映教师真实的教学水平,这在一定程度上影响了教学评价工作.

从以上的分析可以看出,消极评教数据的隐蔽性较强,且受学生行为模式以及不同性质科目的制约较大,因而已有方法难以有效识别. 因此,需要从新的指标维度上建立学生评教数据离群点的检测算法.

学生评教数据的离群点主要有 2 种:(i)学生随意、恶意评教数据;(ii)因学生个人行动模式的不同,趋高或趋低评分数据. 基于以上学生评教数据离

收稿日期:2020-12-17
基金项目:国家自然科学基金(61462037)资助项目.
通信作者:邓 松(1982—),男,江西南昌人,副教授,博士,主要从事实体关联、数据法学和教育信息化研究. E-mail:47817086@qq.com

群点产生的特点,可以看出:

1) 仅从教学质量评价表的总分来判别离群数据是不够的,因为存在第 1 种学生随意评教的总分与积极评教分数相近的可能性.

2) 不能仅考虑学生的评教动机,因为第 2 种离群数据类型的产生并不受学生主观条件的影响.

为有效检测在学生评教数据中的离群点,需要从教师教学质量评价表各项目的得分入手. 教师教学特征的客观映像是一致的,因此不同教师在不同班级科目中存在一个隐含的固定得分结构.

1　相关研究

在学生评教数据中的离群点是与大多数数据点不一致,或是由学生不同的认知结构而产生的数据[3]. Zhi Hongzhi 等[4] 与 Xu Xiaodan 等[5] 全面地分析了离群点的数据类型,将离群点检测手段分为基于邻近、基于分类、基于聚类、基于统计等维度. 吴镜锋等[6] 提出,在异常数据监测实际应用中,异常数据类型通常较少. 可以从无监督或有监督的角度,基于聚类以及邻近方式对数据中的离群点进行检测.

1.1　基于聚类的检测算法

董秋仙等[7] 针对传统聚类算法的准确性问题,通过样本的相异度参数选取聚类中心,有效提高了聚类算法的准确性,并减少了迭代次数,降低了资源的消耗; 王子龙等[8] 为解决聚类算法的局部最优问题,采用维度加权的欧氏距离来度量样本点之间的远近程度,通过样本密度迭代确定初始聚类中心,取得了更好的聚类效果; 张巍等[9] 基于启发式聚类算法,提出了一种系统发育树的随机聚类建树方法,解决了大规模序列数据最优解的问题; 何惠等[10] 提出了基于欧氏空间 k-means 聚类算法的特征向量集,构建对边界集上的半定规划问题,从而达到优化算法模型的目的.

尽管基于聚类的检测算法对离群点较为敏感,且能在保持较高准确性的同时减少资源消耗. 但是学生评教数据通常由多个项目维度构成,基于聚类的检测算法在高维数据中存在时间复杂度过高的问题. 同时评教数据的离群点缺少严格定义,在聚类算法中存在边界模糊问题.

1.2　基于邻近的检测算法

针对高维度数据的离群点问题,仇开等[11] 提出了一种加权 LOF 的异常数据检测算法,在有效解决高维度大数据离群点问题的同时,也具有较低的时间复杂度; 贺寰烨等[12] 提出了一种基于密度空间的局部离群因子算法,将云虚拟机在密度空间中的性质融合到 LOF 算法中. 该算法在保持高检测效率的同时,克服了时间复杂度问题; 梅林等[13] 针对算法在极高维度数据中的时间复杂度与负载均衡问题,提出了一种加权分布式离群点检测方法. 通过基于网格划分的加权分配算法对数据进行优化,使离群点检测算法在极高维度中依旧可以保持较高的性能.

在实际应用中,学生评教数据往往存在大量因学生趋利而分数偏高的数据,使得在数据集中出现 2 个以上的高密度区域. 当趋利评教学生数过多时,传统的算法会出现较高的误检率.

已有的研究在离群点检测效率的提升方面,大多数侧重于资源消耗和时间复杂度的研究,对不同类型数据离群点检测精度的研究还有待进一步深入. 学生评教数据的离群点与常见的风险控制模型不同,虽然现有较多的离群点检测算法,若直接用于学生评教数据的离群点检测,会出现拟合不足的问题. 因此,需要针对学生评教数据特点,设计新的离群点检测模型.

本文基于学生评教数据离群点在数据集中的距离相似度与形状相似度,提出了一种基于几何特征半监督邻近的学生评教数据离群点检测算法,实验表明该方法在学生评教数据中具有较高的识别准确率,且在不同类型学科的评教数据中有较强的适应性.

2　几何特征计算

虽然不同领域数据集的离群点所具有的特征是不同的,但是各领域数据集的几何特征指标是一致的. 本文根据高校教学质量评测表的制作方式和评教数据特点,设计了距离相似度和形状相似度 2 个特征,并给出了特征量化计算方式.

2.1　距离相似度

距离相似度是样本特征图形间对应顶点平均距离的计算指标,主要用于识别学生依据第 2 种方式消极评教产生的离群点. 为便于分析,在 2 维平面上建立样本的特征图形映射,坐标轴表示教学测评表的各项指标(见图 1) .

在图 1 中,特征图形的顶点为样本对应评价指标的得分,连接各顶点构成样本的特征图形. 从图 1 可以看出,离群样本与真实样本之间几何中心重合且形状相近,对应顶点间的距离差别较小. 同时,第 1 种方式消极评教而产生的离群点,存在样本间顶点距离差抵消而误判的可能性,无法直接使用传统

算法进行相似度计算. 因此,需要建立新的算法模型.

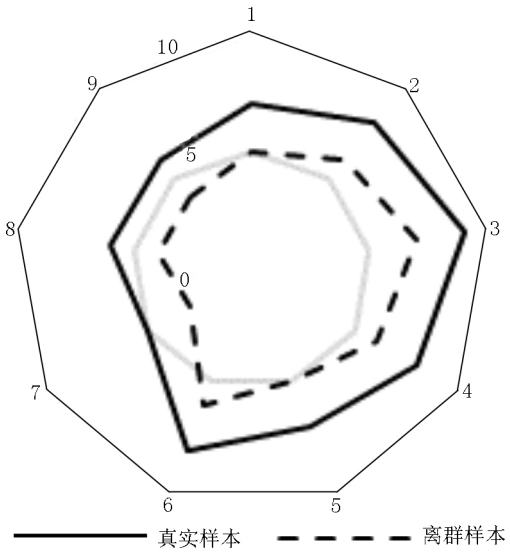


图 1 第 2 种离群样本特征图形

为此,建立 n 维几何特征空间,其中 n 为教学评价表的评价指标数量,空间坐标轴为教学评价表的各评价指标. 根据样本的各指标得分,建立样本在几何特征空间中的点映射.

学生在填写教学评价表时,往往会为避免损害个人与教师之间的利益,选择填写偏高的评价分数,导致样本的映射点存在向空间中最大值方向聚集的趋势. 在数据集中会存在 2 个以上的密度中心,即真实评教数据和偏高的评教数据. 数据呈偏态分布,频数分布的高峰位于分数偏高处,偏度系数小于 0. 此时评教数据中心点受偏高数据的影响,无法体现数据的代表值. 为有效检出偏高和偏低的样本,通过数据集的中值建立超平面 ,对高分样本和低分样本的空间进行分割,坐标轴为评价指标得分,其 2 维效果图如图 2 所示.

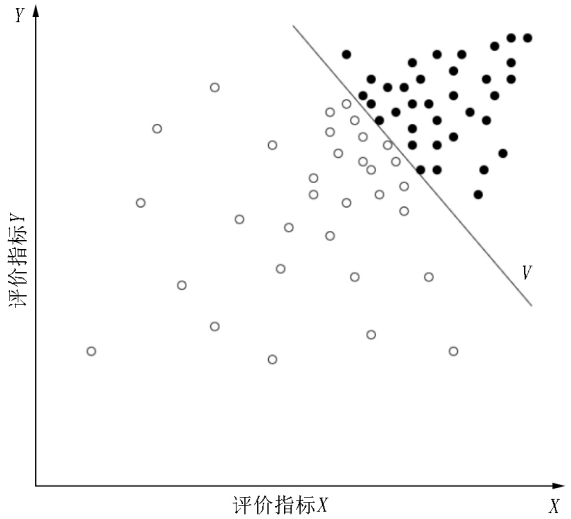


图 2 超平面对样本进行分割

在图 2 中,样本数据集被超平面分割为 2 部分. 其中黑色样本点为高分样本,分值在中值以上; 白色样本点为低分样本,分值在中值以下. 这 2 类样本点到超平面的距离差最小,此时通过原点任意直线上样本点特征图形的形状相同,且对应顶点间的距离差相近. 由于真实样本的产生存在一定的误差,即每个学生个体对教师主观映像的差距,因此无法通过直线对真实样本进行精确拟合. 为便于分析,在判别空间中标记真实样本区域,其 2 维效果图如图 3 所示.

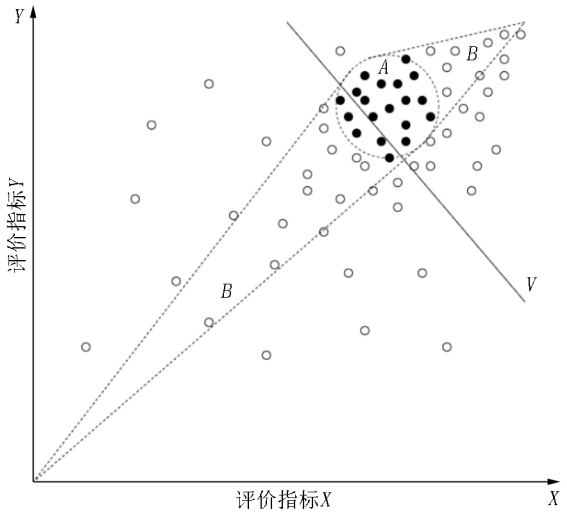


图 3 在判别空间中真实样本的区域

在图 3 中,A 区域为真实样本区域,B 区域为由第 2 种方式消极评教而产生的离群样本区域. 从图 3 可以看出,由第 2 种方式消极评教而产生的离群样本与其对应的真实样本并不是在一条直线上. 其原因是教学测评表的各评教项目存在分数上限,当分数足够高时,样本点会向最大值方向聚集.

为有效计算样本的距离相似度,使用 $l_0' = (\boldsymbol{\omega} x_0 + b) / \| \boldsymbol{\omega} \|$ 计算样本点到超平面 V 的距离,其中 l_0' 为点样本点 x_0 到超平面 V 的距离,$\boldsymbol{\omega}$ 为超平面 V 的法向量. 当 $l_a' = (l_0')_{\max}$ 时,其对应 x_a 的距离相似度相对总体最小,离群度最大,离群值为 l_a'. 在标记检出的样本后,重新计算超平面 V,重复样本距离相似度计算步骤.

2.2 形状相似度

形状相似度为样本间的形状比率,主要用于识别学生依据第 1 种方式消极评教产生的离群点. 为便于分析,在 2 维平面上建立样本的特征图形映射,坐标轴表示教学测评表的各项指标(见图 4).

从图 4 可以看出,离群样本与真实样本的形状比率较小,对应内角差异较大. 此时,离群样本在空间中的映射点分布 2 维效果图如图 5 所示.

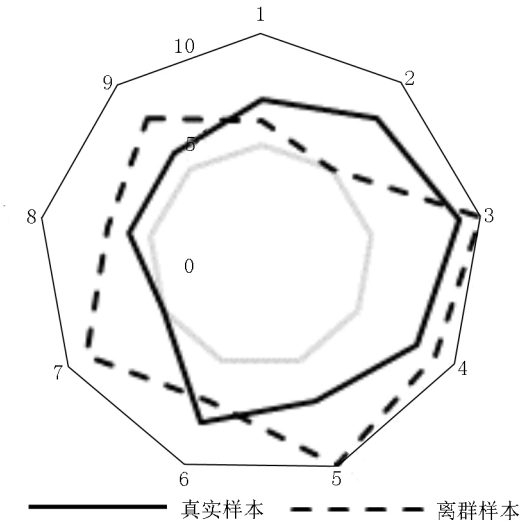


图 4　第 1 种离群样本特征图形

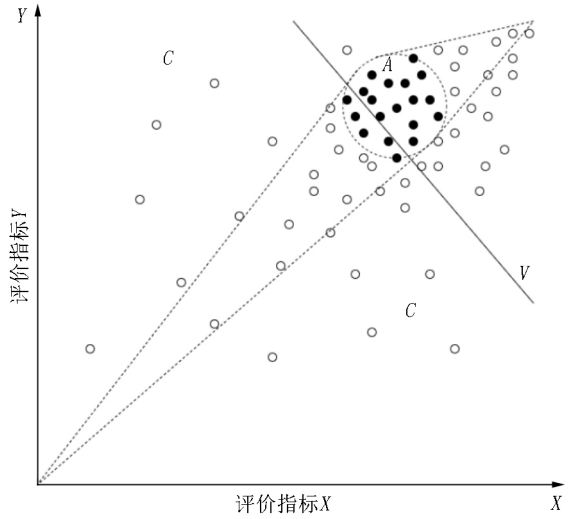


图 5　第 1 种离群样本特征图形

在图5中，V为分割高分样本与低分样本空间的超平面，A 区域为真实数据样本空间，C 区域为依据第 1 种方式产生离群数据的分布空间. 需设计算法有效检出 C 区域中的样本，首先建立数据集的中值在超平面 V 上的法向映射，其 2 维效果图如图 6 所示.

在图 6 中，m 点为数据集中值在超平面上的映射点. 然后建立数据集在超平面上的法向映射点，最后使用 $l''_0 = \sqrt{\sum_{i=1}^{n} (x_i^2 - y_i^2)^2}$ 计算超平面上与中值映射点 m 距离最远的点 o，其中 l''_0 为中值映射点到样本映射点的距离，i 为教学测评表中的评教指标数，x 与 y 为在对应坐标轴上的值. 当 $l''_0 = (l''_0)_{\max}$ 时，样本点离群程度最高，离群值为 l''_a. 在标记检出样本后，重新计算超平面 V，重复样本形状相似度计算步骤.

2.3　离群点检测

为自动化识别学生评教数据中的离群点，本文采用基于半监督近邻的算法. 以距离相似度、形状相似度 2 个分类指标建立判别空间. 当检出的数据中真实数据越少，且检出的样本点与真实数据的样本点距离越大时，算法效果越好.

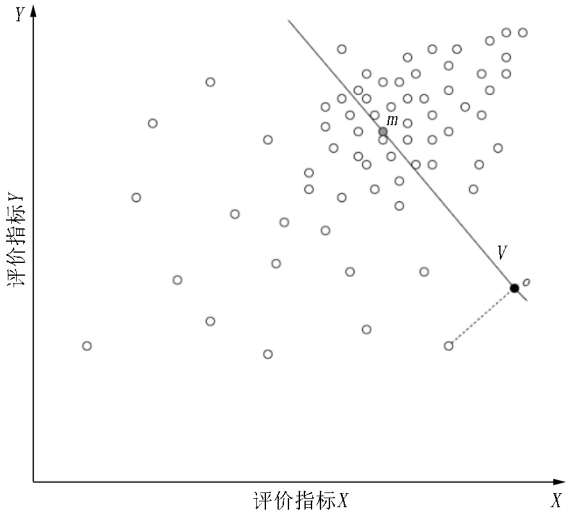


图 6　样本点在超平面上的法向映射

采用距离相似度与形状相似度构建判别空间，分别计算每个样本点的距离相似度和形状相似度，建立样本在判别空间中的点映射. 以样本在判别空间中与原点的距离作为评价依据，依次标记去除判别空间中与原点距离最大的样本，直至达到高校评教工作要求的数据降噪水平. 具体步骤如下:

步骤 1　加载数据，并将数据标准化;

步骤 2　分别计算数据集内样本的距离相似度 T_1、形状相似度 T_2;

步骤 3　计算数据集中样本点的离群程度 K

$$K = \sqrt{(l'_a)^2 + (l''_a)^2}. \tag{1}$$

步骤 4　按照数据点离群程度进行大小排序后存入数据集$\{X_i\}$ 中.

依据以上方法，所有的标记离群点被按顺序存入数据集$\{X_i\}$ 中. 评教工作者可根据实际需要，调整算法的降噪水平.

3　实验分析

本文从江西财经大学教育评价系统中抽选了 2 276 656 条 2018 年、2019 年及 2020 年有效的学生评教数据.

在数据挖掘领域中，一般采用准确率和召回率来评价离群点检测算法的性能，其中准确率是检出的离群点与检出数据的比值，而召回率是检出的离群点与离群数据总数的比值. 但是对于评教工作而言，数据的降噪处于一个固定的水平. 去除的数据量

为高校评教工作所要求的数值,此时召回率受降噪水平的影响,难以体现算法的性能. 同时,因为第 2 种方式产生离群数据的存在,且检出的样本数量 M 取决于高校评教工作的需求,准确率的内涵不再是所选择 M 个样本中包含的真实数据与 M 的比值. 此时准确率的内涵应是检出的 M 个样本在数据集根据样本离群程度降序排列的前 M 个样本中所占的比值,使用 $P = M'/M$ 表示. 其中 P 为算法准确率,M 为算法检测标记的样本数,M' 为数据集中根据样本离群程度降序排列的前 M 个样本与算法检测标记的 M 个样本的交集. 为了更好地评价离群点检测算法的性能,进行以下定义.

定义 1 基于年份的离群点检测算法准确率是指在所选择的 M 个样本中根据数据产生年份(即学生提交教学测评表的年份)对样本进行分类,分别计算每一年度样本集的离群点检测算法准确率.

定义 2 基于学院的离群点检测算法准确率是指根据学生所属学院对所选择的 M 个样本进行分类,分别计算每一学院样本集的离群点检测算法准确率.

定义 3 基于学生性别的离群点检测算法准确率是指在所选择的 M 个样本中依据学生性别对样本集进行分类,分别计算不同性别样本集的离群点检测算法准确率.

定义 4 基于年级的离群点检测算法准确率是指将选择的 M 个样本依据学生的年级进行分类,分别计算每一年级样本集的离群点检测算法准确率.

本文方法先从形状相似度、距离相似度 2 个方面构建判别空间,然后对样本的离群程度进行分析. 实验首先使用本文提出的离群点检测算法与已有方法对评教数据进行离群点检测标记,为保证评教数据的完整性,对离群程度最高的 5% 的数据进行标记;然后比较这 2 种算法之间的检测效果;最后分析样本类型对本文方法的影响.

3.1 基于年份的离群点检测算法准确率

本文的目的是分析计算评教数据中每个样本的离群程度,在实验过程中只把准确率作为评价特征,不考虑召回率. 为了检验本文算法的有效性,以及检测是否存在过拟合问题,即算法是否仅在某一年的样本中有效,或算法在不同年份的数据中是否存在准确率水平差异较大的问题. 进行如下实验:根据定义 1 计算过去 3 年评教数据的离群点检测准确率,并选择相关算法进行对比. 由于本文采用基于半监督邻近的检测算法,因此选取 Zhi Hongzhi 等[14] 使用的 PAM 算法作为对比. 由于 PAM 算法准确率较高,且在不同类型的数据集中有稳定的表现;而且 PAM 算法可以进行基于邻近的方式进行离群程度计算,便于与本文算法进行比较. 实验结果如图 7 所示.

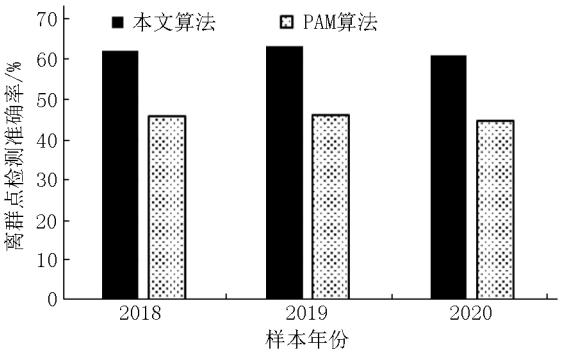


图 7 基于年份的离群点检测算法效果比较

从图 7 可以看出,对于每一年度评教数据离群点检测的准确率,本文方法比 PAM 算法具有明显优势,离群点检测准确率提升了 16.4%. 其原因是:存在部分学生采取第 2 种方式进行评教,脱离教师实际情况直接评出高分,导致学生评教数据产生大量高分的数据簇,使 PAM 算法出现误判,即将在高分处集中的样本判为低离群程度样本. 当算法运行时,基于较高分数段的中心点对低分数段样本进行标记;在低分数段样本标记完毕后,出现将中等分数段样本标记为中等离群程度样本的错误.

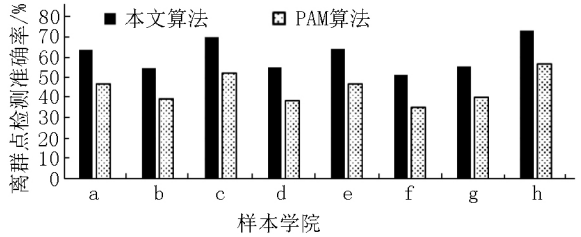
本文算法利用样本数据的几何特征查找样本集中的中间样本,基于样本点间距离进行离群程度计算. 判定点的计算与样本的局部密度无关,不易受到高密度数据簇的影响. 当算法运行时,判定点位于样本的中间位置;在对低分数段样本标记完毕后,会同时对中等分数段和高分数段样本进行评估检测,有效标记高分数段的离群样本. 因此,本文算法的准确率比 PAM 算法具有较大提升.

3.2 基于学院的离群点检测算法准确率

在高校中不同学院的教学模式与教学风格存在一定的差别. 为了检验本文算法在不同学院数据中的有效性,即本文算法在不同教学模式与教学风格下产生的评教数据的有效性. 根据定义 2 计算不同学院评教数据的离群点检测准确率,并使用 PAM 算法进行对比. 实验结果如图 8 所示.

从图 8 可以看出,本文算法的准确率比 PAM 算法更高. 准确率超过 60% 的学院有 4 所,分别是艺术学院、工商管理学院、软件与物联网工程学院、财税与公共管理学院;准确率小于 60% 的学院分别是

外国语学院、人文学院、法学院、统计学院. 由此可以看出，不同学院学科对算法准确率的影响较小. 在图 8 中，本文算法的准确率在不同学院的数据中存在波动，其中统计学院的数据中准确率最低，为 51.3%；艺术学院的数据中准确率最高，为 73.2%. 原因是不同学院学生的评教积极性不同，依据第 2 种方式评出的高分的增多会导致算法的准确率下降，使得算法在不同学院中样本的离群点检测率出现一定的起伏. 由此可知学生的评教积极性可能与教学模式和教学风格有关，会对算法的有效性造成一定的影响.



注：a 为财税与公共管理学院，b 为法学院，c 为工商管理学院，d 为人文学院，e 为软件与物联网工程学院，f 为统计学院，g 为外国语学院，h 为艺术学院.

图 8　基于学院的离群点检测算法效果比较

3.3　基于学生性别的离群点检测算法准确率

为了检验学生性别因素对算法准确率的影响，即学生性别的不同对算法造成的影响，根据定义 3 计算不同性别学生评教数据的离群点检测准确率. 实验结果如图 9 所示.

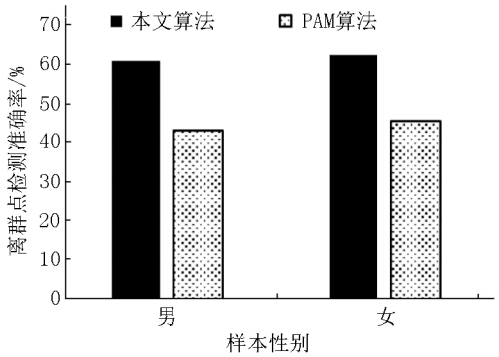


图 9　基于学生性别的离群点检测算法效果比较

从图 9 可以看出，本文算法对不同学生性别的评教数据均保持了较高的准确率，且性别之间算法准确率差异较小. 本文算法在女性学生评教数据中的准确率比男性学生的高出 1.6%，差异较小，这可以认为学生性别对算法的影响较小.

3.4　基于年级的离群点检测算法准确率

为了检验本文算法在不同年级学生中的效果，即学生年级的不同对算法造成的影响，根据定义 4 计算不同年级学生评教数据的离群点检测准确率. 实验结果如图 10 所示.

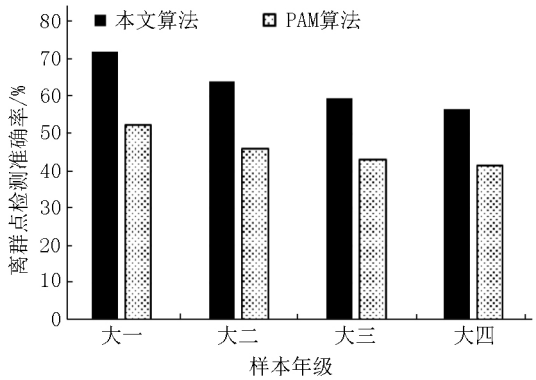


图 10　基于年级的离群点检测算法效果比较

从图 10 可以看出，本文算法在不同年级学生评教数据中准确率均较高. 随着年级的提升，本文算法和 PAM 算法的准确率均出现了下降的趋势. 其原因可能是学生参与评教的积极性随年级的上升而下降；同时任课教师与学生之间的互惠关系会随年级的提升而逐渐累加膨胀[14]. 从而导致根据第 2 种方式随意评教产生的数据逐年增多，使得算法的准确率存在缓慢下降的趋势.

3.5　实验总结

本文利用学生评教数据的几何特征，设计算法分析评教数据的距离相似度和形状相似度，完成离群度的计算. 从数据年份、学院、学生性别、学生年级 4 个方面进行对比实验，实验结果表明本文算法的准确率较传统算法具有较大提升. 原因在于：评教数据中的部分高离群度样本存在较为明显的众数趋势，在判别空间中与真实数据样本距离较近. 此时传统算法容易出现错误，即将高密度且高离群度数据簇误判为低离群度数据；而本文算法不易受到判别空间中高密度数据簇的影响. 当随意评出高分的样本足够多时，本文算法的准确率出现了下降. 原因在于：高分样本数量在数据集中占比的上升，会使数据的判定点向高分处偏移. 当高分样本的占比足够高时，判定点与高分区域距离过近，出现高分区域离群度过低的误判，算法的准确率下降. 若直接去除极高分数的样本，有违教学评价活动的初衷[15]，需要从其他方面减轻其带来的影响.

4　结语

为减少学生评教数据的噪声问题，本文基于评教数据的几何特征设计了一种离群点检测算法. 鉴于评教数据的特性，该算法从距离相似度与形状相似度 2 个方面分别对数据样本点进行离群度计算，

通过建立判别空间对样本离群度进行排序标记,完成学生评教数据中离群数据的检测工作.

实验结果表明:本文方法在提高在学生评教数据中的准确率方面比已有方法具有一定的优势,在高校的教学质量改革中拥有较好的应用前景.在未来的工作中,将进一步研究在学生评教数据离群点去除过程中的极高分数样本问题,以更好满足高校教学评价工作的需求.

5 参考文献

[1] 韩映雄,周林芝. 学生评教的信度、效度、影响因素及应用风险 [J]. 复旦教育论坛,2018,16(6):74-81.

[2] 徐雪珂,邓松,张荣,等. 基于个体特征与教学评价的教师学习对象推荐 [J]. 江西师范大学学报:自然科学版,2019,43(4):409-415.

[3] Aggarwal C C. Outlier analysis [M]. 2nd ed. Cham: Springer,2017:286-286.

[4] Zhi Hongzhi,Bah M J,Hammad M. Progress in outlier detection techniques: a survey [J]. IEEE Access,2019,7:107964-108000.

[5] Xu Xiaodan,Liu Huawen,Yao Minghai. Recent progress of anomaly detection [EB/OL]. [2020-02-19]. https://downloads.hindawi.com/journals/complexity/2019/2686378.pdf.

[6] 吴镜锋,金炜东,唐鹏. 数据异常的监测技术综述 [J]. 计算机科学,2017,44(S2):24-28.

[7] 董秋仙,朱赞生. 一种新的选取初始聚类中心的 k-means 算法 [J]. 统计与决策,2020,36(16):32-35.

[8] 王子龙,李进,宋亚飞. 基于距离和权重改进的 k-means 算法 [J]. 计算机工程与应用,2020,56(23):87-94.

[9] 张巍,王洋,刘东宁,等. 基于随机聚类方法建模的序列分析 [J]. 江西师范大学学报:自然科学版,2017,41(5):470-475.

[10] 何慧,胡小红,覃华,等. 用核 k-means 聚类减样法优化半定规划支持向量机 [J]. 江西师范大学学报:自然科学版,2013,37(6):574-578.

[11] 仇开,姜瑛. 加权 LOF 结合上下文判断的云环境中服务运行数据异常检测方法 [J]. 计算机工程与科学,2020,42(6):951-958.

[12] 贺寰烨,林果园,顾浩,等. 云虚拟机异常检测场景下改进的 LOF 算法 [J]. 计算机工程与应用,2020,56(23):80-86.

[13] 梅林,张凤荔,王瑞锦,等. 基于网格划分加权的分布式离群点检测算法 [J]. 电子科技大学学报,2020,49(6):860-866.

[14] 赵颖,哈巍. 分数膨胀的影响因素与学生选课策略 [J]. 清华大学教育研究,2020,41(6):63-74.

[15] 蒋贵友,郭丽君. 大学评教"共谋"行为及其治理路径 [J]. 大学教育科学,2020(2):105-110.

The Outlier Detection Algorithm for Student Evaluation Data Based on Geometric Features

TANG Yukun,DENG Song*,XU Mengya,GUO Xin

(College of Software and Internet of Things Engineering,Jiangxi University of Finance and Economics,Nanchang Jiangxi 330013,China)

Abstract: To solve the outlier problem in the student evaluation data,a outlier detection algorithm for the students´ evaluation data based on geometric feature is presented according to the way and characteristics of the data that does not fit in with the evaluation data. By analyzing the geometric characteristics of the samples,the algorithm calculates the outlier degree of the samples and completes the outlier detection,which is divided into three steps. Firstly,based on the teaching quality evaluation data,the point mapping of samples is established in the geometric feature space. Secondly,the discriminant space is constructed from the shape similarity and distance similarity. The point mapping of sample points in the discriminant space is obtained by analyzing and calculating the sample points in the geometric feature space. Finally,the samples in the discriminant space are tested based on semi-supervised neighbors. The experimental results show that the algorithm has a high detection accuracy and has a good application value in the teaching effect of university teachers.

Key words: student evaluation; geometric features; outliers; support vector machine

(责任编辑:冉小晓)