MD-KNN算法在高校精准资助中的应用

李 博[1,2],李 霞[1,3],张 晓[1,2],王艳秋[1,2],李 恒[1,2],
张 勇[1,2],凌玉龙[1,2]

(1.西北工业大学 计算机学院,陕西 西安 710129;
2.西北工业大学 工信部大数据存储与管理重点实验室,陕西 西安 710129;
3.西北工业大学 学生资助服务中心,陕西 西安 710129)

摘 要:精准资助是当前一个热点问题,国内很多高校也对学生精准问题进行了深入的探索。为提升高校学生精准资助工作的准确性,采用MD-KNN算法(Mahalanobis distance k-nearest neighbor algorithm)对该问题进行分析。对收集到的数据信息利用基于马氏距离的MD-KNN算法进行聚类,再对聚类结果进行迭代分析,以提高经济困难学生筛选工作的精度。学生群体由于其本身的特殊性,其行为也会与贫困情况有联系,文中对学生行为与贫困情况进行分析:发现学生在学校食堂就餐次数、就餐天数与贫困指数具有正相关的联系。以西安某高校2017年11月至2018年4月学生行为数据为样本进行实验;用生成的名单与线下正常认证的贫困学生名单进行对比。实验证明MD-KNN算法在高校学生精准资助中具有很大的应用价值。

关键词:MD-KNN算法;马氏距离;高校精准资助;聚类算法;数据挖掘
中图分类号:TP311.13 **文献标识码**:A **文章编号**:1673-629X(2020)07-0091-05
doi:10.3969/j.issn.1673-629X.2020.07.020

Application of MD-KNN in Accurate Subsidy of Colleges

LI Bo[1,2],LI Xia[1,3],ZHANG Xiao[1,2],WANG Yan-qiu[1,2],LI Heng[1,2],
ZHANG Yong[1,2],LING Yu-long[1,2]

(1.School of Computer Science,Northwestern Polytechnical University,Xi'an 710129,China;
2.Ministry of Communications Key Laboratory of Big Data Storage and Management,
Northwestern Polytechnical University,Xi'an 710129,China;
3.Student Aid Service Center,Northwestern Polytechnical University,Xi'an 710129,China)

Abstract:Precision subsidy is a hot issue at present. Many colleges in China have explored the problem of students' precision in depth. In order to improve the accuracy of the precise subsidy work for college students,the MD-KNN algorithm (Mahalanobis distance k-nearest neighbor algorithm) is used to analyze the problem. The collected data are clustered by the MD-KNN algorithm based on Mahalanobis distance,and the clustering results are analyzed iteratively,which improves the accuracy of screening for the students whose families are financially difficult. Because of the particularity of students,their behavior will also be related to poverty. We analyze the connection of students' behavior and poverty. It is found that the number of times students eat in school canteens,the number of days they eat have a positive correlation with poverty index. The experiments are based on the data of students' behavior from November 2017 to April 2018 in a university in Xi'an. The results are compared with the list of poverty-stricken students who are normally certified offline. Experiment shows that the MD-KNN algorithm has great application value in the precise subsidy of College students.

Key words:Mahalanobis distance k-nearest neighbor algorithm;Mahalanobis distance;precise subsidy of colleges and universities;clustering algorithm;data mining

0 引 言

学生群体是社会中非常重要的群体,并且对社会的发展有重大影响,因此对学生行为的分析有很大的意义。但是由于学生群体是一个相似度比较高的群

收稿日期:2019-09-24 **修回日期**:2020-01-20
基金项目:国家重点研发计划(2018YFB1004401)
作者简介:李 博(1994-),男,硕士研究生,CCF会员(93705G),研究方向为云存储、数据挖掘;李 霞,博士,研究方向为高校学生资助;张 晓,博士,副教授,研究方向为存储系统。

体,目前针对学生群体的数据挖掘算法还比较少。贫困学生的筛选与资助是很多高校的一项重要事务,通过分析学生的家庭情况、消费和学习行为,可以找到需要资助的贫困学生群体,还可以预防甄别生活规律有异常的学生,从而进行相应的帮助[1-5]。

基于马氏距离的 KNN 算法(Mahalanobis distance k-nearest neighbor algorithm,MD-KNN,马氏 KNN)是一种改进的 KNN 算法。相比于传统的 KNN 算法,MD-KNN 算法采用了马氏距离,可以更好地处理一些非数值型数据,比如:生源地、性别等因素。文中采用该算法,以西安某高校在校学生数据为样本,进行实验分析,探究 MD-KNN 算法在贫困学生资助工作中的效果。在通过 MD-KNN 算法筛选得到拟贫困学生名单后,与实际筛选名单进行对比,分析两者的匹配率,以及学生的消费水平。在分析学生数据时发现:贫困学生的在校就餐次数与就餐天数会有一定的规律,并通过实验分析验证了这一观点。此外还发现,学生吃早餐情况也与该生的学习成绩之间有正相关的联系[6-9]。

1 研究现状

KNN 分类算法是一种经典且应用广泛的数据挖掘算法。随着科学技术的发展,为了适应一些新问题、新背景,在传统 KNN 算法的基础上也不断提出新的改进方法,比如:AHP-KNN(analytic hierarchy process KNN)、FCD-KNN(feature correlation difference KNN)等。MD-KNN 算法是在原先 KNN 算法的基础上,采用马氏距离(Mahalanobis distance)来计算样本之间的距离,因此 MD-KNN 算法可以更多地考虑非数值型因素,从而提升算法的精度。根据在西安某高校收集的学生数据,其中以数值型数据为主,如经济消费数据、学习成绩、图书馆入馆记录、借书记录等,也有部分非数值数据,如生源地、性别、是否残疾单亲等。文中选用 MD-KNN 算法进行学生行为的分析,探索学生行为的规律,并筛选需要资助的学生,以及行为有异常的学生[10-12]。

国内外对于学生群体的行为分析由来已久,20 世纪就有人开始进行研究。随着时代的发展,学生的行为也变得复杂化,但是学生群体内部依然具有较高的相似性。如何对贫困学生进行精确资助,以及分析学生行为,提高学生学习生活质量,保障学生生活安全,成为了各个高校关心的热点问题之一。随着大数据技术的发展,从 2014 年起,很多团队尝试将大数据分析与精准资助相结合,如西北工业大学学生资助服务中心的李霞老师团队。但是现有大部分高校的精准资助系统的算法具有局限性,过于主观,某些高校的贫困学生通过老师或学生人工筛选推荐,缺乏科学的理论分析。文中采用 MD-KNN 算法来进行学生行为的分析,从大数据角度探究学生行为,推动困难学生精准资助领域的发展[13-15]。

2 理论介绍

2.1 MD-KNN 算法介绍

马氏距离是由印度统计学家马哈拉诺比斯(P. C. Mahalanobis)提出的,表示数据的协方差距离[14-15]。经典的 KNN 算法采用的是欧氏距离,欧氏距离单纯地考虑数值上的距离,但是当前在很多的实际场景中需要考虑非数值型的因素,并且很多因素之间并不是相互独立的。马氏距离认为属性之间是存在联系的,比如身高与鞋码之间就是存在联系的,所以在距离计算公式中引入了协方差。而如果是两个完全独立的变量,其协方差是 0,在这种情况下就变成了欧氏距离。对于一个均值为 $\mu = (\mu_1, \mu_2, \cdots, \mu_p)^{\mathrm{T}}$,协方差矩阵为 $\boldsymbol{\Sigma}$ 的多变量向量 $x = (x_1, x_2, \cdots, x_p)^{\mathrm{T}}$,其马氏距离为:

$$D_M(x) = \sqrt{(x-\mu)^{\mathrm{T}} \sum^{-1} (x-\mu)} \tag{1}$$

其中,T 是指矩阵的转置。

在马氏距离的设计中,某一微小变量的作用可以被放大,这在某些应用环境中会导致过度拟合的状况。但不同于其他数据,学生群体是一个具有高相似性的群体,大部分成员内部之间生活作息规律比较相似,就餐时间、地点相对固定且有规律,不同学生样本的行为也是大致相似。而如果通过分析发现一些奇异点,或者某些方面存在异常,则需要学校的额外注意。因此文中根据马氏距离的这一特点,认为采用马氏距离的 MD-KNN 算法更为适合学生数据分析。

2.2 学生在食堂就餐天数与就餐次数的分析

马氏距离考虑了变量之间的相互联系,文中分析这一设计思想,着重分析了贫困学生在学校食堂的就餐次数与就餐天数之间的联系。通常而言,经济困难的在校大学生相比于经济富裕的在校大学生,其娱乐时间和消费水平会较低。而很多大学食堂会有补助,食堂饭菜的价格会略低于学校外饭店的价格。因此,潜在的困难大学生的在校天数和在食堂就餐次数可能更多。文中根据在校学生在学校食堂刷卡产生的消费记录进行分析,列出以下公式:

$$N = (X + Y)/Z \tag{2}$$

其中,X 和 Y 分别表示午餐数和晚餐数,Z 是根据该学生的就餐情况(午餐和晚餐),推断出的该学生在校天数,再乘以 2 得到的数字(该生在食堂应该就餐次数)。最终 N 越大说明该学生在校天数以及食堂就餐

数之间的比例高,该生在校食堂就餐的频率高,也更有可能是经济较为困难的同学。理论上, X 和 Y 可以为不超过在校天数任意大的整数,也可以为 0。文中默认设置每位学生每天只吃一顿午餐(晚餐),即:某位学生在中午时段有多次刷卡记录(比如分开打菜和米饭),文中也会将金额累计,认为是一次消费记录。

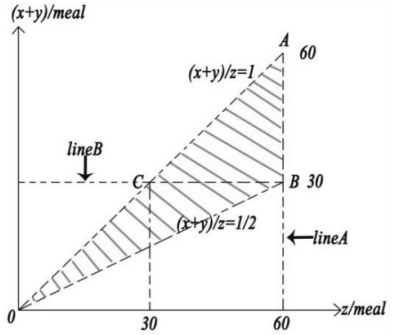

图 1　学生在食堂就餐天数与就餐次数对比

根据式(2),由每个学生在一段时间内的就餐情况得到 N 值(有对应的 $X+Y$, Z)。所有学生消费行为所对应的点,都会落在阴影区域里,即:每个样本点得到 N 值的最大值不会超过 1,最小值不会小于 0.5,($0.5 \leqslant N \leqslant 1$)。如图 1 所示,文中取一个分析区间为 30 天,则应该就餐数目为 60 顿(午餐和晚餐)。可以分析这条线上的 A 、B 两点, A 点是最理想状态,该生在校 30 天,就餐 60 顿, $N=1$。而 B 点,该生就餐 30 次,在校 30 天, $N=0.5$,这名同学的情况很极端,他是每天只吃午餐或晚餐,连续 30 天(比如连续 30 天只吃午餐),则也可以推导出其在校 30 天,但是 $N=0.5$。如果一位同学连续多天均不在食堂消费,则 X 与 Y 都会相应减少,他的数据点会位于该阴影区域的左下角部分,趋向于 0 点。

再沿平行 x 轴方向分析线 B ,线 B 上有两个点, B 和 C ,这两点都是就餐次数为 30 次,但是由于点 B 的行为,他的在校天数是点 C 的两倍(点 C 的在校天数是 15 天)。但是分析推断样本 B 点学生的行为更有规律。通过进一步的分析,推测在学校内消费次数越多和越平均的学生样本更有可能是需要资助的贫困学生。当加入早餐的因素时,图 1 的变化如图 2 所示。

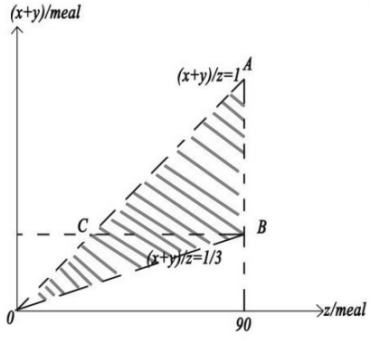

图 2　学生在食堂就餐天数与就餐次数对比(含早餐)

此时 Z 的含义为:根据早中晚餐实际就餐次数推算出来的该生实际在校天数,再乘以 3,得到的该生应当就餐次数。根据分析,经济困难的学生的数据点更有可能落在阴影区域的右上角区域范围(所有学生的数据点都会集中在这个阴影三角形区域),即贫困学生的在校食堂消费次数更多,消费天数更多,消费次数也更均匀、更规律。

3　实验分析

文中搭建 Eclipse+Tomcat 实验环境,使用 Java 语言编程,以西安某高校 2016 和 2017 级硕士研究生,2012 至 2017 级博士研究生在 2017 年 11 月至 2018 年 4 月(约 180 天)的学生行为数据(主要是食堂消费数据、图书馆进出信息、学习成绩等数据)进行实验分析。其中男生 7 636 人(约占 68.36%),女生 3 534 人(约占 31.64%),共计 11 170 人。该高校有 2 个校区,A 校区位于西安大唐西市附近,整体消费水平较高;B 校区位于郊区,物价相对较低;且该高校不同学院位于不同校区。在进行数据分析时,将校区、学院等差异考虑在内。针对所研究的问题,设计了如下三个实验:(1)使用该校实际贫困生名单的实际生活消费数据,对前述学生食堂就餐次数与在校天数的分析进行相应的验证;(2)使用 MD-KNN 算法,对该高校学生进行贫困学生的筛选,然后比较与已有的,由人工认定贫困学生的名单的差异;(3)为了更好地对比实验(2)和人工认定贫困学生的名单,将这两份名单中的学生进行经济消费水平的对比。

3.1　关于经济困难学生在食堂就餐次数、就餐天数的分析

通过分析,经济困难学生会更多地在校内食堂就餐,因此其校内食堂就餐次数与就餐天数会相对较高,获得资助的同学其数据实验结果会落在图 1 所示三角区域的右上角部分。文中通过采集西安某高校人工认定的 200 多名贫困学生在 2017 年 11 月至 2018 年 4 月间,学校食堂的早餐、午餐、晚餐的就餐情况的数据进行验证,结果如图 3 所示。

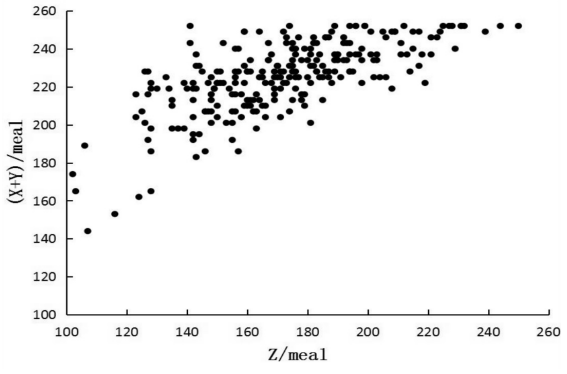

图 3　贫困学生在食堂就餐天数与就餐次数对比图

如图 3 所示,该校人工筛选的贫困学生就餐情况是符合文中分析的,学生在食堂的就餐次数与就餐天数两种因素与学生的贫困与否是一种正相关的条件,贫困学生相比于非贫困学生会在学校食堂就餐次数更多,这也为今后贫困学生资助工作提供了一种新的参考因素。

3.2 MD-KNN 算法的实验分析

根据收集到的实验样本数据,使用 MD-KNN 算法进行分析,设置经济、消费、学习、生源地、是否有生源地贫困证明、是否残疾等二十余项标签,然后进行迭代的实验分析。在得到初步的贫困学生名单后,再在结果中设置筛选学生名单条件,即:拟评选人数、助学金等级等,这样就得到了由 MD-KNN 算法筛选推荐的贫困学生名单,筛选出的部分学生名单见表 1。将由 MD-KNN 算法筛选得到的名单与实际人工审核推荐的学生名单进行对比。两份名单的匹配率大致在 50% 左右,这一概率并不算高,但分析原因可能有两方面:(1)使用的 MD-KNN 算法或许还需要进行改进,以更好适应高校贫困学生筛选的应用环境;(2)人工筛选名单具有很大的不确定性,老师、学生很多情况下是通过申请表、平时的认知(甚至并不认识)来进行筛选推荐,人工筛选贫困学生也存在一些漏洞。因此通过实验 3.3,对两份名单中的学生进行消费情况的分析。

表 1 MD-KNN 算法筛选得到的经济困难学生名单(部分)

编号	性别	学院名称	年级	月均消费	日均消费	就餐次数	在校天数	补助等级
1	女	生命学院	2017	463.5	15.93	205	83	二级补助
2	男	软件与微电子学院	2017	425.84	17.38	140	73	二级补助
3	男	理学院	2017	385.1	15.67	165	75	一级补助
4	男	计算机学院	2017	430.34	17.01	188	83	一级补助
5	女	计算机学院	2017	331.29	14.07	156	80	二级补助
6	女	计算机学院	2017	339.47	13.39	186	82	二级补助
7	男	计算机学院	2017	440.64	15.06	204	75	一级补助
8	女	计算机学院	2017	336.54	13.97	160	76	二级补助
9	女	计算机学院	2017	315.88	12.75	178	78	一级补助
10	男	计算机学院	2017	384.47	15.44	157	80	一级补助

3.3 MD-KNN 算法与线下人工筛选名单的对比

针对 3.2 节实验分析的结果,对两份名单中的学生进行进一步的分析。还是以 2017 年 11 月至 2018 年 4 月之间的学生消费数据来进行对比,实验结果如图 4 所示。

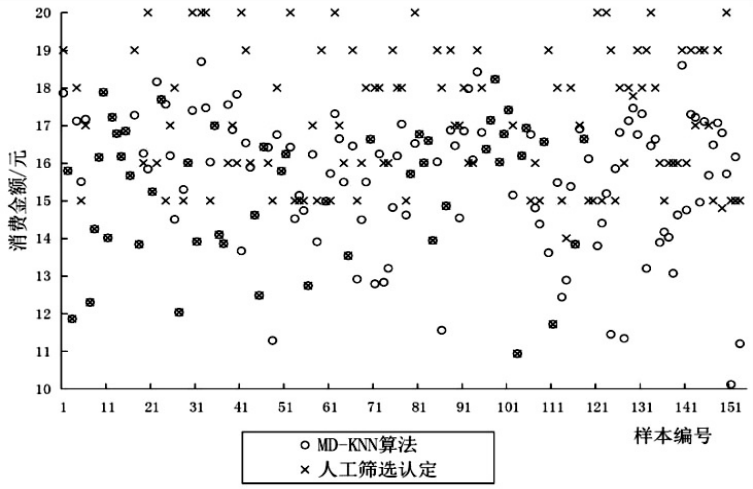

图 4 日均消费金额对比

通过对比发现,由文中筛选的学生名单的消费水平(图 4)明显低于由学校提供的,由实际人员参与评审所得到的贫困学生名单,这就说明所设计的贫困学生筛选算法是有效的。虽然一些经济困难学生由于身体或疾病原因可能会有较高的消费数据,但总体而言,大部分经济困难的学生在学生群体中的消费数据应该

是较低的。因此,MD-KNN 算法在筛选困难学生的过程中是一种有效的算法,并值得进一步的分析研究。

4　结束语

学生群体是一个相似度较高的群体,具有很多共性,对学生群体进行行为分析,筛选出应该资助的贫困学生,是当前很多高校的一项重要事务。通过分析 MD-KNN 的特性,将其应用到贫困学生筛选资助的过程中,设置学生的属性标签、消费行为标签、学习行为标签(相同条件下最后考虑学习成绩)进行筛选,发现与实际得到的贫困学生名单相比,通过 MD-KNN 算法筛选出来的学生名单消费水平更低,有更高的精确度。同时发现,经济水平较低的学生的在校食堂消费天数与消费次数更高,以及学习成绩与吃早餐次数具有正相关的关系。因此该研究是有效的,有助于贫困学生资助工作的发展。

参考文献:

[1] GUO Jinyu, WANG Xin, LI Yuan. KNN based on probability density for fault detection in multimodal processes[J]. Journal of Chemometrics, 2018, 32(7): e3021.

[2] TEKA A, BAIRAGI S, SHAHADAT M, et al. Poly(vinylidene fluoride) (PVDF)/potassium sodium niobate (KNN)-based nanofibrous web: a unique nanogenerator for renewable energy harvesting and investigating the role of KNN nanostructures[J]. Polymers for Advanced Technologies, 2018, 29(9): 2537-2544.

[3] 职为梅,张　婷,范　明. 基于影响函数的 k-近邻分类[J]. 电子与信息学报,2015,37(7):1626-1632.

[4] 宓文斌. 数据挖掘在银行信贷业务中的应用[D]. 上海:上海交通大学,2012.

[5] LI Bo, ZHANG Xiao, YAN Jingyi. Design of intelligent shopping cart for supermarket[J]. Journal of Physics: Conference Series, 2019, 1207(1): 010025.

[6] 李　博,张　晓,颜靖艺,等. 基于值差度量和聚类优化的 K 最近邻算法在银行客户行为预测中的应用[J]. 计算机应用,2019,39(9):2784-2788.

[7] SEETHA H, SARAVANAN R, MURTY M N. Pattern synthesis using multiple kernel learning for efficient SVM classification[J]. Cybernetics and Information Technologies, 2012, 12(4): 77-94.

[8] TAHERI S, MAMMADOV M. Learning the naive Bayes classifier with optimization models[J]. International Journal of Applied Mathematics and Computer Science, 2013, 23(4): 787-795.

[9] CHEN Xi, ZHONG Wenqi, WANG Tiancai, et al. Genetic optimization of energy consumption of pellet shaft furnace combustor based on support vector machine (SVM)[J]. International Journal of Chemical Reactor Engineering, 2014, 12(1): 205-214.

[10] 路敦利,宁　芊,臧　军. 基于 BP 神经网络决策的 KNN 改进算法[J]. 计算机应用,2017,37:65-67.

[11] 李正杰,黄　刚. 基于 Hadoop 平台的 SVM_KNN 分类算法的研究[J]. 计算机技术与发展,2016,26(3):75-79.

[12] FEKI-SAHNOUN W, NJAH H, HAMZA A, et al. Using general linear model, Bayesian networks and Naive Bayes classifier for prediction of Karenia selliformis occurrences and blooms[J]. Ecological Informatics, 2018, 43: 12-23.

[13] SAINI I, SINGH D, KHOSLA A. QRS detection using K-nearest neighbor algorithm (KNN) and evaluation on standard ECG databases[J]. Journal of Advanced Research, 2013, 4(4): 331-344.

[14] 林　钢,季　薇. 基于迭代决策树的帕金森 UPDRS 预测模型研究[J]. 计算机技术与发展,2019,29(1):216-220.

[15] 高茂庭,段元波. 结合用户聚类和评分偏好的推荐算法[J]. 计算机应用研究,2018,35(8):2260-2264.