

文章编号: 2095-0365(2018)02-0102-05

基于聚类的博士学位论文质量分析

刘能现, 姜云飞

(福州大学 研究生院 福建 福州 350116)

摘 要: 博士学位论文是衡量博士研究生培养质量的重要指标。本文阐述了数据挖掘在教育领域的应用意义及聚类分析技术,通过对影响博士学位论文质量的相关属性进行分析处理,基于微软 SSAS 平台构建了博士学位论文质量聚类分析挖掘模型,以福州大学博士学位论文相关数据为实例进行聚类分析,找出影响博士学位论文质量的主要因素,从生源质量、科研条件、指导教师、在学期间发表论文、奖学金资助等方面为提高博士学位论文质量提供参考依据。

关键词: 聚类算法; 博士研究生; 学位论文质量

中图分类号: G40-057 **文献标识码:** A **DOI:** 10.13319/j.cnki.sjztdxxbskb.2018.02.18

一、引言

博士研究生教育作为国民教育的顶端,其教育的质量和数量是衡量一个国家高等教育水平的重要指标,一流的博士研究生教育是“双一流”建设的重要内容。随着我国经济社会发展进入新常态,对提高博士研究生教育质量提出了新任务。2017年1月国务院学位委员会第三十三次会议上刘延东副总理指出质量是研究生教育的生命线,要加强质量保障体系建设,强化培养单位质量意识,抓好质量监控,着力提高博士生培养质量。博士学位论文是博士生获得博士学位时最具代表性的研究成果,是衡量博士生能否毕业和授予学位的一个重要依据,是博士生学习阶段学术水平与研究能力的集中体现,是衡量博士生培养质量的重要指标^[1]。

数据挖掘技术可以从大量的数据中发现隐藏的模式与知识,目前已成功应用在金融、生物医学和电子商务等广泛的领域。近年来,在教育信

息化、远程教育及在线学习系统等应用的带动下,教育数据挖掘开始受到越来越多的研究者的关注^[2]。教育数据挖掘是指应用数据挖掘方法从教育数据中提取出有意义的信息的过程,在大数据时代,对教育数据进行挖掘分析可以帮助发现和解决教育中的各类问题,如为管理人员提供辅助决策依据、帮助教师改进课程、提升教学效果以及提高学生的学习效率等。

本文基于数据挖掘中的聚类算法,对福州大学2015—2017学年毕业的306位博士研究生的学位论文相关数据进行聚类分析与挖掘,找出影响博士学位论文质量的主要因素,为提高博士学位论文质量提供参考依据。

二、聚类分析技术

聚类分析^[3]是一种重要的数据分析方法,就是将数据对象分成类或簇的过程,使同一个类中的对象之间具有较高的相似性,而不同类中的对象具有较大的差异性。聚类分析数据挖掘过程可

收稿日期: 2017-10-22

基金项目: 福建省中青年教育科研项目(GJAS170084)“基于数据挖掘技术的研究生生源质量研究”;福州大学教育管理基金资助项目(16SKZ23)“数据挖掘技术在研究生培养质量评估中的应用”

作者简介: 刘能现(1983-),男,助理研究员,研究方向: 研究生教育。

本文信息: 刘能现,姜云飞. 基于聚类的博士学位论文质量分析[J]. 石家庄铁道大学学报: 社会科学版, 2018, 12(2): 102-106.

概括为三部分: 数据准备、聚类分析及结果的解释和评估。

目前, 聚类分析的算法^[4-5]大致可分为层次聚类算法、划分式聚类算法、基于密度和网格的聚类算法和其他聚类算法。在众多的聚类算法中, K 均值聚类(K -means)是最经典、最方便快捷的算法之一。

K 均值聚类算法^[6]的过程概述如下:

(1) 从 N 个样本数据中随机选取 K 个对象作为初始的聚类中心;

(2) 分别计算每个样本到各个聚类中心的距离, 将对象分配到距离最近的聚类中;

(3) 所有对象分配完成后, 重新计算 K 个聚类中心;

(4) 与前一次计算得到的 K 个聚类中心比较, 若聚类中心发生变化, 转(2), 否则转(5);

(5) 当聚类中心不发生变化时, 停止并输出聚类结果。

三、聚类算法在博士学位论文质量分析中的应用

(一) 数据的选取及预处理

福州大学的研究生教育管理系统记录了研究生从招生入学到毕业授学位的全过程的培养信息, 为有效进行博士学位论文数据聚类分析提供了基础。该系统数据库主要包括学生基本信息、导师信息、课程信息、学习成绩信息、奖学金信息、开题情况、中期考核、发表论文登记信息、论文送审以及答辩成绩等数据表。但是这些数据表涉及很多属性, 哪些字段属性被用于聚类算法, 直接影响到最终的聚类效果。本文目标是对博士学位论文质量进行聚类分析, 经查阅相关文献^[7-8], 论文质量的主要影响因素包括博士生招生政策、博士生生源质量、导师学术水平、学科建设水平、科研条件、培养管理机制、学位论文的评审与答辩制度等。因此, 选取了学号、攻读方式、学习方式、导师称谓、学科、课程成绩、获奖情况、论文选题来源、发表论文情况、论文送审成绩、答辩成绩等字段作为挖掘的属性。

为了让数据满足挖掘要求和提高聚类算法的效率, 对数据使用下列规则进行预处理。

(1) 学生基本信息处理, 将属性攻读方式分为公开招考、提前攻博和硕博连读, 将属性学习方

式分为全日制学习和非全日制学习, 将属性导师称谓分为国家级人才(如院士、长江学者、国家杰出青年基金获得者等)和非国家级人才。

(2) 课程学习成绩处理, 每个学生根据研究方向的不同可能选修不一样的课程和学分, 为公平衡量学生学习成绩, 采用学位课加权成绩作为学生综合成绩, 即学生综合成绩等于所有学位课程的成绩乘以课程学分之和除以所有学位课程的学分之和。为更好地进行聚类, 采用等频分箱法进一步对成绩进行离散化处理, 即将计算的成绩按从大到小进行排序, 并将学生成绩等分为四部分, 分别标记为 A、B、C 和 D 四个等级。

(3) 奖学金信息处理, 分为有获奖学金和未获奖学金并分别标记为是和否。

(4) 论文选题来源信息处理, 博士学位论文选题可能来自导师不同级别的科研项目或自选课题, 为便于聚类把论文选题来源分为国家级项目、省部级项目、其他项目和其他。

(5) 发表论文情况, 根据学校博士生在读期间发表论文要求规定, 博士生至少要发表 1 篇一类论文才能申请毕业, 因此根据博士生在读期间发表一类期刊论文数量将发表论文情况分为 4 个等级, 发表 1 篇一类期刊论文的标记为 1, 发表 2 篇一类期刊论文的标记为 2, 发表 3 篇一类期刊论文的标记为 3, 发表 4 篇及以上一类期刊论文的标记为 4。

(6) 学位论文信息处理, 根据论文送审和答辩成绩确定博士学位论文质量等级, 首先根据论文答辩成绩把学位论文等级分为优、良、中、差 4 个等级, 根据实际情况, 论文答辩成绩一般会好于送审成绩, 考虑到论文送审采用匿名盲审方式, 送审成绩能比较客观地体现论文质量, 因此, 把送审成绩出现及格及以下而答辩成绩为中及以上的论文等级定为差。

(二) 建立博士学位论文质量聚类分析挖掘模型

利用 Microsoft SQL Server Analysis Services (SSAS) 数据挖掘平台作为实验平台。在该平台上建立博士学位论文质量聚类分析挖掘模型, 主要包括以下步骤: (1) 创建 Analysis Services 项目; (2) 创建数据源; (3) 创建数据源视图; (4) 创建挖掘结构模型和设置算法参数, 本文选择聚类分析中 K -means 算法, 经过多次反复实

验,当参数聚类数设为4时,能得到具有明显特征的数据聚类结果;(5)部署和处理挖掘模型。在

Visual Studio 2008 中创建的挖掘模型如图1所示。

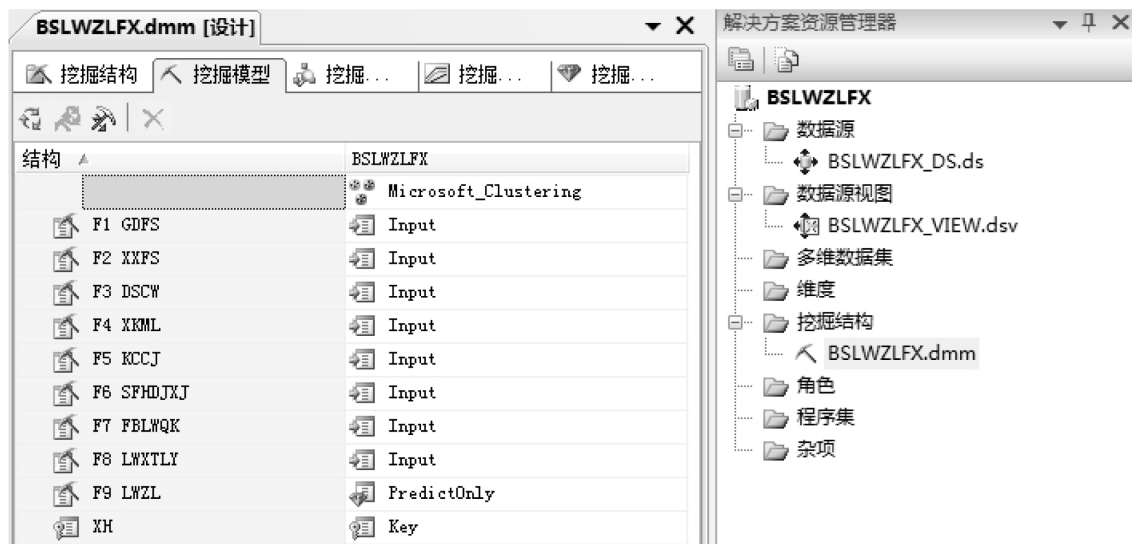


图1 博士学位论文质量聚类分析挖掘示意图

(三) 结果的解释和评估

对福州大学2015至2017学年毕业的306位博士研究生的学位论文相关数据,执行博士学位论文质量聚类分析挖掘模型,得到聚类分析结果,如表1所示。从聚类结果可以看出,数据被划分为4类:

第一类包括103名学生,占总数的33.7%,该类主要特征包括攻读方式以公开招考为主,学习方式以非全日制为主,导师中国国家级人才占0.9%,学科门类中理学和工学占61.6%,其他学科占38.4%,课程成绩处于A、B等级的占38.4%,课程成绩处于C、D等级的占61.6%,在学期间只有0.9%的学生获得了奖学金,在学期间58.9%的学生只发表了1篇一类期刊论文,发表论文情况较一般,论文选题有43.8%获得国家级项目支持,论文选题有28.6%未获得项目支持,论文质量优良率为62.5%。总得来说,该类学生在学期间表现相对较差,主要原因包括该类学生中大部分为非全日制学生、有较多的学位论文未获得项目的支持和法学等非优势学科占比较大。

第二类包括109名研究生,占总数的35.6%,该类主要特征包括攻读方式以提前攻博和硕博连读为主,学习方式以全日制为主,导师中国国家级人才占18.9%,学科门类中理学和工学占94.4%,其他学科占5.6%,课程成绩处于A、B等级的占

58.5%,课程成绩处于C、D等级的占41.5%,在学期间有85.8%的学生获得了奖学金,在学期间65.1%的学生发表了3篇以上一类期刊论文,发表论文情况较好,论文选题有89.6%获得国家级项目支持,论文选题只有3.8%未获得项目支持,论文质量优良率为90.6%,其中优占73.6%。总得来说,该类学生在学期间表现较为优秀,主要得益于该类中大部分学生为提前攻博和硕博连读的优秀全日制学生、有较多的学位论文获得国家级项目的支持和理学等优势学科占比较大。

第三类包括62名研究生,占总数的20.3%,该类主要特征包括攻读方式以公开招考为主,学习方式全日制占57.4%,非全日制占42.6%,导师中国国家级人才占1.9%,学科门类中理学和工学占81.5%,其他学科占18.5%,课程成绩处于A、B等级的占46.3%,课程成绩处于C、D等级的占53.7%,在学期间有11.1%的学生获得了奖学金,在学期间64.8%的学生发表了2篇一类期刊论文,发表论文情况较一般,论文选题有22.2%获得国家级项目支持,有33.3%获得省级项目支持,有24.1%未获得项目支持,论文质量优良率为66.6%。总得来说,该类学生在学期间表现相对一般。

第四类包括32名研究生,占总数的10.4%,该类主要特征包括攻读方式以公开招考为主,学习方式以全日制为主,导师中国国家级人才占

77.4% ,学科门类中理学和工学占 71.1% ,其他学科占 29% ,课程成绩处于 A、B 等级的占 74.2% ,课程成绩处于 C、D 等级的占 35.8% ,在学期间有 74.2% 的学生获得了奖学金 ,在学期间 42% 的学生发表了 3 篇以上一类期刊论文 ,发表论文情况较好 ,论文选题有 87.1% 获得国家级项目支持 ,

论文选题只有 6.5% 未获得项目支持 ,论文质量优良率为 90.3% ,其中优占 61.3% 。总得来说 ,该类学生在学期间表现较好 ,虽然该类学生以公开招考的普通全日制学生为主 ,但该类中导师为国家级人才比例及学位论文获国家级项目支持比例都较高 ,所以该类学生表现也较好。

表1 数据聚类分析结果统计表

一级指标	二级指标	总体 (306)	分类 1 (103) / %	分类 2 (109) / %	分类 3 (62) / %	分类 4 (32) / %
攻读方式	公开招考	194	98.2	8.5	83.3	90.3
	提前攻博	69	0.9	54.7	13.0	6.5
	硕博连读	43	0.9	36.8	3.7	3.2
学习方式	全日制	171	10.7	100.0	57.4	64.5
	非全日制	135	89.3	0.0	42.6	35.5
导师称谓	非国家级人才	260	99.1	81.1	98.1	22.6
	国家级人才	46	0.9	18.9	1.9	77.4
学科门类	工学	132	50.9	23.6	70.4	35.5
	理学	105	10.7	70.8	11.1	35.5
	管理学	53	27.7	2.8	16.7	29.0
	法学	11	8.9	0.0	1.9	0.0
	医学	5	1.8	2.8	0.0	0.0
课程成绩	A	76	18.8	28.3	14.8	51.6
	B	79	19.6	30.2	31.5	22.6
	C	79	33.9	24.5	18.5	12.9
	D	73	27.7	17.0	35.2	12.9
是否获得奖学金	否	184	99.1	14.2	88.9	25.8
	是	122	0.9	85.8	11.1	74.2
发表论文情况	1	96	58.9	19.8	3.7	19.4
	2	71	6.3	15.1	64.8	38.7
	3	58	18.8	22.6	18.5	6.5
	4	82	16.1	42.5	13.0	35.5
论文选题来源	国家级项目	185	43.8	89.6	22.2	87.1
	省部级项目	45	19.6	4.7	33.3	0.0
	其他项目	24	8.0	1.9	20.4	6.5
	其他	52	28.6	3.8	24.1	6.5
论文质量	优	136	16.1	73.6	37.0	61.3
	良	96	46.4	17.0	29.6	29.0
	中	54	25.0	7.5	25.9	9.7
	差	20	12.5	1.9	7.4	0.0

根据聚类分析结果 ,可以得出以下结论及建议:

(1) 生源质量是影响学位论文质量的源头 ,学校应深化招生体制改革 ,提高全日制博士生的比例 ,特别是提高硕博连读和提前攻博学生的比例 ,逐渐减少在职人员的招生比例 ,同时学校应加

强学科建设 ,提高学校吸引力 ,千方百计提高生源质量。

(2) 在学期间参与科研项目是影响学位论文质量的重要因素 ,参与科研项目能够锻炼博士生科研能力。因此 ,要鼓励博士生在学期间多参与

科研项目,同时学校在博士研究生招生指标分配时应向有国家级科研项目的导师倾斜。

(3) 导师学术水平是产出高质量学位论文的重要因素,学校应加强博士生导师队伍建设,深化博士生导师遴选和聘任上岗改革,学校在博士研究生招生指标分配时应向学术水平较高的导师倾斜,同时对考核不合格的导师进行停招处理。

(4) 在学期间发表论文情况和博士学位论文质量是正相关的,因此,要鼓励博士研究生在学期间积极撰写发表高质量的学术论文,为撰写学位论文打下坚实的基础,从而有效地提高学位论文质量。

(5) 奖学金和博士学位论文质量也是正相关的,学校应加大对博士研究生的奖助力度,减轻他们的生活经济压力,发挥奖学金激励作用,不断调动博士研究生的研究积极性和激发他们的科研创新能力,从而促进学位论文质量进一步提高。

四、结论

博士学位论文是衡量博士研究生培养质量的重要指标,博士学位论文质量受生源质量、导师学术水平、科研条件等诸多因素影响。本文通过对福州大学博士研究生的学位论文相关数据进行聚类分析,探讨博士生招生政策、博士生生源质量、导师学术水平、学科建设水平、科研条件、培养管理机制、学位论文的评审与答辩制度等因素与博士学位论文质量的关联关系。聚类结果表明,博士学位论文质量主要与攻读方式有关,其他因素也有一定关系,但作用较小。从提高生源质量、鼓励多参与科研项目、加强博士生导师队伍建设、鼓励在学期间发表高质量论文、加大奖学金资助力度等方面为学校提高博士学位论文质量及博士研究生培养质量提供参考依据。

参考文献:

- [1]李艳,马陆亭,赵世奎.博士学位论文质量及其影响因素研究[J].江苏高教,2015(2):105-109.
- [2]周庆,牟超,杨丹.教育数据挖掘研究进展综述[J].软件学报,2015,11(26):3026-3042.
- [3]韩家炜.数据挖掘概念与技术[M].北京:机械工业出版社,2007.
- [4]孙吉贵,刘杰,赵连宇.聚类算法研究[J].软件学报,2008(1):48-61.
- [5]姜云飞.基于聚类的博士生选拔方法探析[J].长春大学学报,2014,24(2):238-240.
- [6]张良均,樊哲,赵云龙,等.Hadoop大数据分析与应用实战[M].北京:机械工业出版社,2017:84.
- [7]孔令夷.基于解释结构模型的博士学位论文质量关键影响因素分析[J].中国高教研究,2012(4):51-55.
- [8]郭巍,郑舒婷.博士学位论文质量影响因素及保障体系建设[J].沈阳师范大学学报,2014,38(3):146-148.

Analysis of Doctoral Dissertation Quality Based on Clustering

Liu Nengxian, Jiang Yunfei

(Graduate School, Fuzhou University, Fuzhou 350116, China)

Abstract: Doctoral dissertation is an important index to evaluate the quality of doctoral graduate education. This paper describes the significance of data mining in education and the clustering analysis technology. By analyzing and processing the affecting factors of doctoral dissertation quality, a cluster analysis mining model of doctoral dissertation quality based on Microsoft SSAS platform is established. This paper takes the doctoral dissertation data of Fuzhou University as an example to make a cluster analysis, and then finds out the main factors affecting the quality of the doctoral dissertation, which provides a reference for improving the quality of doctoral dissertation from the aspects of the quality of students, research conditions, guidance tutors, publishing papers during the academic period, and scholarships etc.

Key words: clustering algorithm; doctoral candidate; dissertation quality