doi:10.13582/j.cnki.1674-5884.2019.03.016

大学生就业影响因素中基于粗糙集的智能数据分析方法

柳媛慧[1],陈林书[2],马庆[3]

(1.湖南科技大学 外国语学院,湖南 湘潭 411201;2.湖南科技大学 计算机科学与工程学院,湖南 湘潭 411201;
3.湖南软件职业学院 软件与信息工程学院,湖南 湘潭 411201)

摘　要:大学生就业形势越来越严峻,影响就业的因素众多,各因素对就业情况的影响并不相同,且相互之间存在关联性。基于粗糙集理论,提出大学生就业因素重要度的定量度量方法,建立基于粗糙集的智能数据分析模型,实验结果表明了新型方法的有效性,能够为大学生的在校学习和就业规划提供了重要指导,并为高校培养和企业招聘优秀大学生提供了决策支持。

关键词:大学生就业;粗糙集;属性重要度;智能分析

中图分类号:G647　　**文献标志码:**A　　**文章编号:**1672-7835(2019)03-0083-05

近年来,随着高校毕业生数量的增加,大学生就业情况变得十分严峻,学生就业压力越来越大[1]。大学生就业的影响因素很多,对就业的重要性各不相同,并且影响因素背后隐含着大量相关的、不完整的和有效的数据[2]。因此,非常有必要对大学生就业影响因素进行分析,找出影响高校学生就业情况的关键因素,进一步挖掘这些关键因素和就业情况之间的依赖关系,找出社会需要的应用型人才,进而为大学生的在校学习和就业规划提供重要指导,并为高校培养和企业招聘优秀大学生提供决策支持[3-4]。

传统的大学生就业影响因素分析方法是数理统计方法,即对平均值、标准差、可信度和覆盖率等指标进行统计和分析,但它最大的缺点是需要根据专家知识库提供主观经验值(比如上述指标的参考值)进行对比,容易由经验值的主观性带来误差[5-6]。粗糙集理论,正好能够弥补数理统计方法的这一劣势,它具有成熟数学基础、不需要任何具有主观性的先验知识,完全根据已有知识库进行分析,挖掘潜伏期的隐含知识和规则,具有客观性和价值性。因此,粗糙集与数理统计方法之间具有很强的互补性[7-10]。

粗糙集理论,由波兰数学家 Pawlak 于 1982 年提出[11],是一种强大的数据分析理论工具,其主要思想是保持信息系统分类能力不变的前提下,通过知识约简导出问题的分类或决策规则,从而发掘潜在的规律和知识[12]。目前,粗糙集理论成功地在数据挖掘、决策分析、机器学习等领域都获得了广泛应用[13-15]。

本文借助粗糙集理论这一强大的数学工具,研究大学生就业因素中基于粗糙集的智能数据分析方法,先介绍了决策表、上近似、下近似、边界域和参数重要度等粗糙集理论基础知识,然后利用粗糙集的属性重要度这一重要概念,提出了大学生就业因素重要度的定量度量方法,部署了实验过程和结果分析,表明上述方法的有效性。

1　粗糙集理论基础

粗糙集模型中的属性重要度是大学生就业因素中基于粗糙集的智能数据分析方法的理论基础,本节将给出粗糙集理论的决策表、下近似、上近似、边界域和属性重要度等概念的形式化定义

收稿日期:20181114

基金项目:湖南省自然科学基金项目(2018JJ2131);湖南科技大学潇湘学院 2018 年教学改革研究项目(x905-G31895)

作者简介:柳媛慧(1982-),女,湖南浏阳人,硕士,讲师,主要从事英语教学与信息化研究。

及其简要描述。

定义 1(决策表)称四元组 $K=(U,S,V,f)$ 是一个信息系统或知识库,简单记为 $K=(U,S)$,其中论域 U 是对象的有限集合,属性 S 是非空有限集合,属性值域 $V=\cup_{a\in A}V_a$,V_a 表示属性 $a\in A$ 的值域,信息函数 $f:U\times A\to V$ 是一个映射。进一步地,称 K 为一个决策表,若满足 S 中属性可划分为两个不相交的子集:条件属性集合 C 和决策属性集合 D,其中 $S=C\cup D$ 且 $C\cap D\neq\varphi$,记为 $K=(U,C\cup D)$。特别地,称 K 为一个单决策表,若满足 $D=\{d\}$,记为 $K=(U,C\cup\{d\})$。

在实际应用中,通常用一张二维表表示信息系统,其中行表示研究对象,列表示对象属性,属性值表示对象信息,一个列属性对应一个等价关系,一个二维表对应一族等价关系。

决策表是一类特殊且非常重要的信息系统,多数决策问题在具体应用中都可以用决策表来解决。

定义 2(下近似与上近似)给定知识库 $K=(U,S)$,$\forall X\subseteq U$ 和一个等价关系 $R\in S$,则定义 R 相对 X 的下近似和上近似分别为:

$$\underline{RX}=\cup\{Y\mid(\forall Y\in U/R)\wedge(Y\subseteq X)\},\tag{1}$$

$$\overline{RX}=\cup\{Y\mid(\forall Y\in U/R)\wedge(Y\cap X\neq\varphi)\}。\tag{2}$$

集合 $pos_R=\underline{RX}$ 称为 R 相对 X 的正域,它表示那些根据知识 R 判断肯定属于 X 的论域 U 对象构成的集合;集合 $neg_R=U-\overline{RX}$ 称为 R 相对 X 的负域,它表示那些根据知识 R 判断不肯定属于 X 的论域 U 对象构成的集合;集合 $bn_RX=\overline{RX}-\underline{RX}$ 称为 R 相对 X 的边界域,它表示那些根据知识 R 即不能判断肯定属于 X 又不能判断肯定不属于 X 的论域 U 对象构成的集合。

定义 3(属性重要度)给定知识库 $K=(U,C\cup D)$,$r\in C$,令划分 $U/D=\{X_1,X_2,\cdots,X_n\}$,则称 $sig(r,D)$ 为条件 r 相对决策 D 的属性重要度,简称为 r 的属性重要度,其中 $|U|$ 为集合 U 的基数,$sig(r,D)$ 如下所示:

$$sig(r,D)=\frac{\sum_{i=1}^{n}|U-bn_rX_i|}{n|U|}\tag{3}$$

显然,有 $sig(r,D)\geqslant 0$ 时,且其值越大,说明该属性的重要度越大,反之该属性的重要度越小,特别地,当 $sig(r,D)=0$ 时,说明属性 r 是不重要的,是可以删除的。

2　实验方法与结果分析

本节先详细交代了实验条件、实验数据选择、实验数据离散化方法等实验数据准备工作;接着,以粗糙集的属性重要度为理论依据,结合实验过程阐述了大学生就业因素的智能数据分析方法;最后,对实验结果进行了分析和总结。

2.1　实验数据准备

在本文的实验中,为了讨论方便,我们选取了家庭背景、英语过级、自身素质、所学专业、就业意愿等就业因素作为条件属性,以就业情况作为单决策属性,并以随机的 15 个大学生信息作为论域,具体信息如下:

(1)论域 $U=\{1,2,3,4,5,6,7,8,9,11,12,13,14,15\}$。

(2)属性集 $C=\{a,b,c,d,e\}$,分别表示属性:a,家庭背景;b,自身素质;c,英语过级;d,所学专业;e,就业意愿。

(3)决策集为 $D=\{f\}$,f 表示就业情况。

为便于分析就业因素的重要度,需要进一步将每个就业因素的取值进行离散化,具体按如下方式进行离散化。

$C(a)$ 家庭背景:1,强;2,一般;3,弱。

$C(b)$ 自身素质:1,优秀;2,良好;3,一般;4,较差。

$C(c)$ 英语过级:1,专业水平;2,六级;3,四级;4,四级以下。

$C(d)$ 所学专业:1,很热;2,热;3,冷;4,很冷。

$C(e)$ 就业意愿:1,强;2,一般。

$D(f)$ 就业情况:1,就业;2,待业。

随机选取了 15 个大学生的就业信息作为数据记录,根据上述就业因素及其离散化结果,可得到部分毕业生信息的原始数据构成的决策表,如表 1 所示。

2.2　基于属性重要度的就业因素分析方法

基于上述实验数据,以粗糙集理论中的属性重要度为重要理论依据,本节讨论基于属性重要度的大学生就业因素分析方法。

根据定义 3 中属性重要度的计算公式,可分别

求出表 1 中的条件属性 C(a 家庭背景、b 自身素质、c 英语过级、d 所学专业、e 就业意愿)相对决策属性 D(就业情况)的属性重要度,如下所示:

$U=\{1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15\}$,$C=\{a, b, c, d, e\}$,$D=\{f\}$,

$U/D=\{\{1, 8, 9, 12, 14, 15\}, \{2, 3, 4, 5, 6, 7, 10, 11, 13\}\}=\{X_1, X_2\}$,

则根据定义 3 可知,决策 D 的划分 $U/D=\{X_1, X_2, \cdots, X_n\}=\{X_1, X_2\}$,其中 $n=2$,$X_1=\{1, 8, 9, 12, 14, 15\}$,$X_2=\{2, 3, 4, 5, 6, 7, 10, 11, 13\}$。

表 1　部分毕业生信息的决策表

ID	属性 C					决策 D
	a	b	c	d	e	f
1	2	3	1	2	2	1
2	2	2	2	4	2	2
3	3	2	2	2	2	2
4	2	2	3	2	2	2
5	2	2	3	4	2	2
6	3	2	3	1	1	2
7	3	3	3	3	1	2
8	2	1	2	2	2	1
9	2	1	4	2	2	1
10	3	2	4	2	1	2
11	2	3	3	1	1	2
12	2	1	3	3	2	1
13	2	2	3	3	1	2
14	3	2	1	2	2	1
15	1	2	4	2	2	1

下面分别计算就业因素 a、b、c、d 和 e 的属性重要度。

(1)计算条件属性中家庭背景 a 相对决策属性 D 的属性重要度如下:

先计算家庭背景 a 的划分,

$U/a=\{\{1, 2, 4, 5, 8, 9, 11, 12, 13\}, \{3, 6, 7, 10, 14\}, \{15\}\}$,

根据定义 2,则 a 相对 X_1 和 X_2 的下近似、上近似和边界域分别如下,

$\underline{aX_1}=\{15\}$,$\overline{aX_1}=U$,$bn_aX_1=\overline{aX_1}-\underline{aX_1}=U-\{15\}$,

$\underline{aX_2}=\varphi$,$\overline{aX_2}=U-\{15\}$,$bn_aX_2=\overline{aX_2}-\underline{aX_2}=U-\{15\}$,

则根据定义 3,可求得 a 相对 D 的属性重要度为

$$sig(a,D)=\frac{\sum_{i=1}^{n}|U-bn_RX_i|}{n|U|}=\frac{|U-bn_aX_1|+|U-bn_aX_2|}{n|U|}=\frac{1+1}{2\times 15}=\frac{2}{30}。$$

(2)同理,可计算条件属性中自身素质 b 相对决策属性 D 的属性重要度如下:

先计算自身素质 b 的划分,

$U/b=\{1, 7, 9\}, \{2, 3, 4, 5, 6, 10, 13, 14, 15\}, \{\{8, 9, 12\}\}$,

根据定义 2,则 b 相对 X_1 和 X_2 的边界域为:$bn_bX_1=bn_bX_2=U-\{8,9,12\}$,

则根据定义 3,可求得 b 相对 D 的属性重要度为 $sig(b,D)=\frac{3+3}{2\times 15}=\frac{6}{30}$。

(3)同理,可计算条件属性中英语过级 c 相对决策属性 D 的属性重要度如下:

先计算英语过级 c 的划分,

$U/c=\{\{1, 14\}, \{2, 3, 8\}, \{4, 5, 6, 7, 11, 12, 13\}, \{9, 10, 15\}\}$,

根据定义 2,则 c 相对 X_1 和 X_2 的边界域为:$bn_cX_1=bn_cX_2=U-\{1,14\}$,

则根据定义 3,可求得 c 相对 D 的属性重要度为 $sig(c,D)=\frac{2+2}{2\times 15}=\frac{4}{30}$。

(4)同理,可计算条件属性中所学专业 d 相对决策属性 D 的属性重要度如下:

先计算所学专业 d 的划分,

$U/d=\{\{1, 3, 4, 8, 9, 10, 14, 15\}, \{2, 5\}, \{6, 11\}, \{7, 12, 13\}\}$,

根据定义 2,则 d 相对 X_1 和 X_2 的边界域为:$bn_dX_1=bn_dX_2=U-\{2,5,6,11\}$,

则根据定义 3,可求得 d 相对 D 的属性重要度为 $sig(d,C,D)=\frac{4+4}{2\times 15}=\frac{8}{30}$。

(5)同理,可计算条件属性中就业意愿 e 相对决策属性 D 的属性重要度如下:

先计算就业意愿 e 的划分,

$U/e=\{\{1, 2, 3, 4, 5, 8, 9, 12, 14, 15\}, \{6, 7, 10, 11, 13\}\}$,

根据定义 2,则 e 相对 X_1 和 X_2 的边界域为:$bn_eX_1=bn_eX_2=U-\{6,7,10,11,13\}$,

则根据定义 3，可求得 e 相对 D 的属性重要度为 $sig(e,D)=\frac{5+5}{2\times 15}=\frac{10}{30}$。

2.3　结果分析

由上述计算结果可知，条件属性 a、b、c、d 和 e 的属性重要度的序关系为 $sig(e,D) > sig(d,D) > sig(b,D) > sig(c,D) > sig(a,D)$，也就是说，在所有大学生就业因素中，对决策因素（就业情况）的重要性程度从高到低分别为就业意愿、所学专业、自身素质、英语水平和家庭背景。根据上面的结果，可以求得大学生就业因素中的家庭背景、英语过级、自身素质、所学专业和就业意愿对就业情况的影响程度的饼形图，如图 1 所示。容易看出，大学生自己较强的就业意愿对就业情况最重要，占 33.3%；所学专业对就业情况的重要性次之，占 26.7%；而相对来说，大学生的家庭背景对就业情况的影响最小，仅为 6.7%。

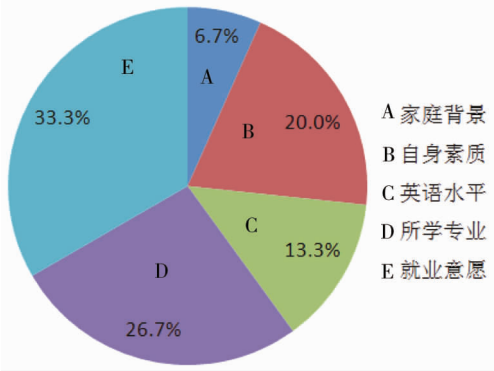

图 1　大学生就业因素对就业的影响权重

事实上，在当前严峻的就业形势下，真正对大学生就业起指导作用的就业因素很多，远远不止家庭背景、英语过级、自身素质、所学专业和就业意愿这 5 个因素，比如还有性别、年龄、身高、计算机水平、籍贯、是否党员等。同时，在实际的大学生就业系统中，很多数据记录的属性值具有相同值、空数据、不一致、冗余存储等噪声现象，这就需要我们对部分数据进行预处理。

因此，在本次实验中，我们采用了强大的粗糙集数据分析工具集 Rosetta 2.0，它具有数据导入导出、补全、离散化、知识约简、过滤、分类规则生成、等价类和上下近似集获取等功能。

3　结语

本文是在已有研究[16-20]的基础上，将粗糙集理论应用于大学生就业因素分析过程，利用粗糙集的属性重要度这一重要概念，提出了大学生就业因素重要度的定量度量方法，定量计算出了家庭背景、英语过级、自身素质、所学专业、就业意愿等就业因素对就业情况的重要性程度，从而建立了大学生就业因素中基于粗糙集的智能数据分析模型。实验结果表明：新型方法准确地求出了就业因素的重要性程度，为大学生的在校学习和就业规划提供了重要指导，并为高校培养和企业招聘优秀大学生提供了决策支持。

参考文献：

[1] 张常新.大学生就业质量的影响因素及对策[J].继续教育研究，2016(1)：113-114.

[2] 张斌，蒋怀滨，王叶飞，等.大学生就业焦虑影响因素及对策研究[J].教育教学论坛，2015(28)：3-4.

[3] 刘洋.高校大学生就业困难的因素与对策分析[J].淮北职业技术学院学报，2016(2)：32-33.

[4] 陆广峰.高校大学生就业难因素分析及对策研究[J].中国成人教育，2015(14)：57-59.

[5] Macaro E. Strategies for language learning and for language use：Revising the theoretical framework[J]. The Modern Language Journal，2006 (3)：320-337.

[6] 周章明，曾鸿鹄，王敦球.基于全面质量观的大学生就业质量评估指标体系构建探析[J].当代教育理论与实践，2012(11)：43-45.

[7] 段明秀.基于 PSO 优化的模糊 RBF 神经网络学习算法及其应用[J].当代教育理论与实践，2010(1)：101-104.

[8] 唐贡如，吴莉.独立学院大学英语教改探析——多媒体教学与基于多元智能理论差异教学的结合[J].当代教育理论与实践，2014(11)：110-111.

[9] 邹新月，周志强.产业结构视角下大学生就业困境与对策研究[J].当代教育理论与实践，2010(5)：98-101.

[10] 唐启涛，张燕，彭利红.基于粗糙集约简算法的配置文本聚类方法研究[J].计算机技术与发展，2015(11)：105-109.

[11] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Science，1982(5)：341-256

[12] 苗夺谦，李道国.粗糙集理论、算法与应用[M].北京：清华大学出版社，2008.

[13] 王国胤.Rough 集理论与知识获取[M].西安：西安交通大学出版社，2001.

[14] Pickering M J，Garrod S. Toward a mechanistic psychology of dialogue[J]. Behavioral & Brain Sciences，

2004 (2):169-190.
[15] 张文修,吴伟志,梁吉业,等.粗糙集理论与方法[M].北京:科学出版社,2001.
[16] Chen L S, Wang J Y (eds.). The Models of Granular System and Algebraic Quotient Space in Granular Computing[J]. Chinese Journal of Electronics, 2016 (6): 1109-1113.
[17] Chen L S, Wang J Y (eds.). Quotient space model based on algebraic structure[J]. High Technology Letters, 2016 (2): 160-169.
[18] Chen L S, Wang J Y. The rough representation and measurement of quotient structure in algebraic quotient space model[J]. High Technology Letters, 2017 (3): 293-297.
[19] Chen L S, Wang J Y, Li L. A new granular computing model based on algebraic structure[J]. Chinese Journal of Electronics, 2019(1): 136-142.
[20] 陈林书,柳媛慧.P2P 网络中基于节点能力自适应的搜索算法[J].湖南科技大学学报(自然科学版),2009(2):61-65.

Analytical Method of Intelligent Data Based on Rough Set Theory in Influence Factors of College Students' Employment

LIU Yuanhui[a], CHEN Linshu[b], MA Qing[c]
(a. School of Foreign Studies; b. School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China; c. School of Software and Information Engineering, Hunan Software Vocational Institute, Xiangtan 411201, China)

Abstract: The situation of college students' employment is increasingly serious. There are numerous factors influencing college students' employment. However, their influences are different, which are connected to each other. Based on rough set theory, a quantitative measurement method of the importance of college students' employment factors, is put forward, and an intelligent data analyzing model of college students' employment factors based on rough set is also established. The experimental results show that the new methods are effective, which provides important guidance for college students' studying on campus and employment planning, as well as offers decision supports for university workers cultivating and corporation recruiting excellent college students.

Key words: college students' employment; rough set; attribute importance; intelligent analysis

(责任校对　龙四清)