基于粗糙集的大学生学习与就业关系分析

吴汉卿[1],吴缓缓[1],杨莹莹[1],纪 霞[1,2]
(1. 安徽大学 计算机科学与技术学院,安徽 合肥 230601;
2. 安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039)

摘 要:随着社会经济的高速发展,人才需求的多元化,国内的教育事业进入了大众化的新时期,毕业生数量逐年增加,导致了高校毕业生就业形势越来越严峻,学生就业难已成为当前社会的热点问题。大学生在校学习成绩作为学生智力、学习态度等因素的直观结果,较为准确地反映了学生的整体水平,也与学生就业有着紧密的联系。为了帮助高校学生合理利用在校学习时间,有导向的进行学习,采集了已毕业计算机科学与技术专业学生在校学习成绩和就业信息数据,利用邻域粗糙集的基本理论,对预处理后的学生成绩就业信息表中的课程属性进行约简,并对得出的属性约简子集进行了详细分析,将学习与就业之间的比较准确的内在联系提供给在校学生,帮助学生找到心仪合适的工作。

关键词:邻域粗糙集;属性约简;学习成绩;就业情况
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2017)05-0188-04
doi:10.3969/j.issn.1673-629X.2017.05.039

Relationship Analysis of Undergraduate Students' School Records versus Employment Based on Rough Set

WU Han-qing[1], WU Huan-huan[1], YANG Ying-ying[1], JI Xia[1,2]
(1. College of Computer Science and Technology, Anhui University, Heifei 230601, China;
2. Key Lab of IC&SP of Ministry of Education, Anhui University, Heifei 230039, China)

Abstract: With the rapid development of social economy, the diversification of demand for talent, the domestic education has entered a new era of popularization. The number of graduates has increased year by year, resulting in increasingly severe employment situation of college graduates, therefore the student employment has become a hot issue of society. As an intuitive result of students' intelligence, learning attitude and other factors, students' academic performance more accurately reflects the overall level of the students, and also has a close relationship with employment. In order to help college students make reasonable use of their time for guiding learning, the achievement and employment data of the graduate majored in computer science and technology have been collected with basic theory of neighborhood rough set to reduce the grade attribute of student learning-employment table that has been preprocessed. Analysis on attribute reduction subset obtained has been carried out which could provide a more accurately intrinsic link between learning and employment to help undergraduate students find a favorite job.

Key words: neighborhood rough set; attribute reduction; academic record; employment situation

0 引 言

一直以来大学生学习与就业都是热点话题,尤其近年来高校扩招,毕业生数量大幅提升,导致就业压力逐年增加。而且,在就业过程中由于对就业准备不充分、缺乏有力的指导以及就业信息不完整,也使在校大学生错失了大量的就业机会。

粗糙集理论[1-3]是 Pawlak 教授于 1982 年提出的一种能够处理模糊和不确定知识的数学工具。粗糙集理论可以在保持决策能力不变的条件下,对属性进行约简,从而发现潜在的知识和规律。该理论最显著的优势在于,在处理不确定和不精确问题时,无需提供其他先验信息。其基本思想是通过关系数据库分类归纳形成概念和规则,通过等价关系的分类以及分类对于目标的近似实现知识发现[4]。

收稿日期:2016-05-09 **修回日期**:2016-08-17 **网络出版时间**:2017-03-13
基金项目:国家自然科学基金资助项目(61402005);安徽省自然科学基金项目(1508085MF127,1308085QF114);安徽大学创新训练项目(201510357190);计算智能与信号处理教育部重点实验室课题项目
作者简介:吴汉卿(1994-),男,研究方向为软件工程;纪 霞,博士,讲师,研究方向为不精确信息处理、粗糙集理论等。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20170313.1545.006.html

为了帮助高校学生合理利用在校学习时间,找到合适的工作,首先收集了已毕业学生的在校学习成绩及就业信息,并进行了相应的预处理。然后利用邻域粗糙集[5-6]的基本属性约简算法进行属性约简,挖掘学生就业与在校学习成绩之间的内在联系。实验结果表明,学生应该着重掌握专业所开设的主干课程,对主干课程的熟练掌握有助于提高学生的综合素质,有利于学生就业[7-9]。

1　邻域粗糙集理论

1.1　基本概念

定义 1:给定一个非空集合 $U=\{x_1,x_2,\cdots,x_n\}$,存在一个距离度量函数满足:

(1)非负性:$f(x_i,x_j)\geqslant 0$,如果 $x_i=x_j$,则 $f(x_i,x_j)=0$。

(2)对称性:$f(x_i,x_j)=f(x_j,x_i)$。

(3)三角不等关系:$f(x_i,x_j)\leqslant f(x_i,x_k)+f(x_k,x_j)$。

定义 2:假设 $<U,f>$ 是一个容忍空间,$\forall x_i\in U,\theta\geqslant 0$,有 $\theta(x_i)=\{x\mid f(x_i,x)\leqslant\theta,x\in U\}$,即 $\theta(x_i)$ 是 x_i 的 θ 邻域集。

定义 3:在邻域决策系统 NDS= $<U,C\cup D,\theta>$ 中,B 是 C 的一个子集,对于任意的 $X\subseteq U$,X 在 B 上的下近似和上近似可分别表示为:

$$\underline{N_B}X=\{x_i\mid\theta(x_i)\subseteq X,x_i\in U\}$$

$$\overline{N_B}X=\{x_i\mid\theta(x_i)\cap X\neq\varnothing,x_i\in U\}$$

其中 $\theta(x_i)$ 定义如下:

$$\theta(x_i)=\{x\mid f(B(x_i),B(x))\leqslant\theta,x\in U\}$$

$B(x)$ 是一个萃取函数,用于提取记录中相应属性的值。

定义 4:在领域决策系统 NDS= $<U,C\cup D,\theta>$ 中,$D_1,D_2,\cdots,D_n$ 是相应决策值的子集。决策属性 D 关于条件属性 B 的下近似和上近似分别为:

$$\underline{N_B}D=\bigcup_{i=1}^{n}\underline{N_B}D_i=\mathrm{POS}_B(D)$$

$$\overline{N_B}C=\bigcup_{i=1}^{n}\overline{N_B}D_i$$

$$\underline{N_B}D_i=\{x_i\mid\theta_B(x_i)\subseteq D_i,x_i\in U\}$$

$$\overline{N_B}D_i=\{x_i\mid\theta_B(x_i)\cap D_i\neq\varnothing,x_i\in U\}$$

决策属性的下近似又被称为决策属性的正域,记为 $\mathrm{POS}_B(D)$。

定义 5:在领域决策系统 NDS= $<U,C\cup D,\theta>$ 中,B 是 C 的一个子集,D 对 B 的依赖度记为:

$$\gamma_B(D)=\frac{|\mathrm{POS}_B(D)|}{|U|}$$

显然,依赖度的大小 $\gamma_B(D)\in[0,1]$。

定义 6:在领域决策系统 NDS= $<U,C\cup D,\theta>$ 中,B 属于 C,B 是一个相对约简,当

(1) $\gamma_B(D)=\gamma_C(D)$;

(2) $\forall a\in B,\gamma_{B-a}<\gamma_B(D)$。

1.2　属性约简算法

根据属性重要度指标,可以构造贪心式属性约简算法[10]。初始化属性约简子集为空,循环计算剩余属性的重要度,选择重要度最大的属性加入约简子集中,直至所有剩余属性的重要度为 0,约简算法终止[11]。

具体的属性约简算法使用了 Liu 等[12]提出的一种优化算法。算法使用了一种基于哈希的方法,将记录集划分到一系列的桶中,并且证明了,每个记录的 θ 邻域,只可能存在于当前桶,以及相邻桶中。该方法将算法复杂度降低到了 $O(m^2|U|)$,其中 m 为属性的个数,$|U|$ 为记录的个数。

(1)改进的快速正域求解算法。

输入:U,P,D,θ;

输出:$F=\{F_1,F_2,\cdots,F_{|u|}\}$。

Step1:将 F 中的每个元素置为 0。

Step2:将 U 划分到对应的桶中。

Step3:对于 U 中每个记录 $x_i(x_i\in B_k)$,判断集合 $B_{k-1}\cup B_k\cup B_{k+1}$ 中的每个记录 x_i 是否存在 $f(P(x_i),P(x_j))\leqslant\theta$ and $\mathrm{Decision}(x_i)\neq\mathrm{Decision}(x_j)$。若存在,则 x_i 不属于当前所求正域;若否,则 x_i 属于当前所有正域,并将对应的 F_i 置为 1。

(2)快速属性约简算法。

输入:U,C,D,θ;

输出:reduce(属性约简结果)。

Step1:reduce 置空。

Step2:当 U 不为空,则执行 Step3,否则执行 Step4。

Step3:对于 $C-\mathrm{reduce}$ 中的每个属性 a,求解 $\mathrm{reduce}\cup\{a\}$ 的正域大小,并选取使正域集合最大的 a 及其对应的正域集合。若得到的最大正域集合不为空,则 reduce=reduce $\cup\{a\}$,并且从记录集 U 中删除最大正域集合中对应的记录。继续执行 Step3。

Step4:结束。

2　数据预处理

2.1　成绩预处理

实验共收集了某大学计算机科学与技术专业 192 名大学生的成绩,共分为 12 种就业类型(见表 1)。由于学校对素质教育较为重视,对于在校学生,除本专业必修、选修课程之外,还需选修相应学分的素质课程和跨专业课程,导致所收集到的数据在某些属性(成绩)上较为稀疏。在实验之前,对数据进行如下处理:

Step1:依次选取每一属性。

表 1 就业类型

序号	就业类型	序号	就业类型
1	其他专业技术人员	7	经济业务人员
2	升学	8	待就业
3	工程技术人员	9	文学艺术工作人员
4	其他人员	10	科学研究人员
5	办事人员和有关人员	11	商业和服务业人员
6	金融业务人员	12	教学人员

Step2:判断在当前属性上非空记录的个数是否大于总记录的一半。

Step3:若是,则保留该属性,并将空记录设置为当前属性的平均值。

Step4:若否,去除该属性。

依据以上思路,共保留 75 个条件属性(课程成绩),1 个决策属性(就业类型)。

预处理前后学生成绩就业信息表分别见表 2 和表 3。

2.2 邻域半径选取

邻域粗糙集中,邻域的大小 θ 作为关键参数,它的选取将直接影响属性约简的结果。文献[5]中使用固定 θ 值($\theta=0.125$)作为所有数据集上属性约简的邻域半径。文献[13]中使用标准差作为 θ 值,即将每一列的属性值做标准差之后,再将这些标准差取标准差作为邻域半径 θ。之所以采用文献[13]提出的邻域半径选取方法,有两点原因。其一,不同数据集描述的物体不同,不存在某一固定的 θ 值作为邻域半径,若采用文献[5]中 θ 值对成绩就业数据进行属性约简,极易产生误差。其二,标准差的方法能够反映数据在平均值上波动的大小。

表 2 预处理前学生成绩就业信息表(部分)

记录编号	常用软件	高级语言程序设计	电子与通信工程概论	常用软件实验	高等数学 A(一)	高级语言程序设计实验	大学生职业生涯规划	…	就业类型
1	2.8	3.2	3.4	3.4	3.6	4	4	…	1
2	4	3.6		2.7	4	3.6	4	…	2
5	3.2	3.6		3.4	1.3	4	4	…	3
10	3.2	4		3.4	3.6	4	4	…	4
15	3.6	1.8		3.4	3.6	2.8	4	…	5
16	2.3	3.2		3.4	1.3	3.6	3.4	…	6
18	3.2	3.6		3.4	2.3	4	4	…	7
23	2.8	3.2		3.4	1.8	2.7	4	…	8
24	3.2	1.3		3.4	1.3	3.4	4	…	9
26	2.8	2.3		3.4	2.8	3.6	4	…	10
107	1.8	3.2		3.4	4	2.8	3.4	…	11
145	2.8	3.2		3.4	2.3	4	4	…	12

表 3 预处理后学生成绩就业信息表(部分)

记录编号	常用软件	高级语言程序设计	常用软件实验	高等数学 A(一)	高级语言程序设计实验	大学生职业生涯规划	大学英语(一)	…	就业类型
1	2.8	3.2	3.4	3.6	4	4	2.8	…	1
2	4	3.6	2.7	4	3.6	4	3.2	…	2
5	3.2	3.6	3.4	1.3	4	4	3.2	…	3
10	3.2	4	3.4	3.6	4	4	2.8	…	4
15	3.6	1.8	3.4	3.6	2.8	4	3.2	…	5
16	2.3	3.2	3.4	1.3	3.6	3.4	1.8	…	6
18	3.2	3.6	3.4	2.3	4	4	2.8	…	7
23	2.8	3.2	3.4	1.8	2.7	4	2.8	…	8
24	3.2	1.3	3.4	1.3	3.4	4	3.6	…	9
26	2.8	2.3	3.4	2.8	3.6	4	3.6	…	10
107	1.8	3.2	3.4	4	2.8	3.4	1.3	…	11
145	2.8	3.2	3.4	2.3	4	4	3.2	…	12

3 实验分析

根据文献[3]中提出的快速属性约简算法,对成绩就业数据进行属性约简,得出属性约简子集对应的课程分别为:数据库原理、高等数学(一)、电路原理、操作系统、大学英语(一)。

从约简结果来看,数据库原理课程对于就业的影响最为重要,这有两点原因:首先,数据库原理是计算机专业课程体系中的高阶课程,课程的学习需要大量其他专业课程的理论基础。其次,这也是对工程应用的真实反映,因为大多数的软件都应用了各种各样的数据库,熟练掌握数据库理论及应用已成为工程技术人员的必备技能。电路原理作为专业基础课程,是学习计算机组成原理的先修课程。操作系统作为计算机专业的核心课程,其先修课程包括高级语言程序设计、数据结构、计算机组成原理等,所以操作系统的课程成绩不仅是对本课程学习情况的概括,也反映了其先修课程的学习情况,而且也是编译原理、数据库原理等课程的重要基础。从计算机专业课程体系来讲,电路原理、操作系统、数据库原理分别作为计算机专业的基础课程、中阶课程和高阶课程[14],三门课程的成绩是其他课程学习情况的综合反映,是对学生专业知识学习水平的高度概括。而高等数学、英语作为公共基础课程,为计算机专业课程的学习奠定了基础,对学生的整体水平有很大的提升。

4　结束语

为了帮助高校学生有导向的进行学习,打好扎实的就业基础,采用了邻域粗糙集上的一种快速属性约简算法,对预处理后的成绩就业数据进行属性约简,得出了学生在校学习情况与就业的内在联系,即学生应该着重掌握本专业的基础课程、中阶课程和高阶课程,对于当前数据集而言,分别对应着电路原理、操作系统、数据库原理三门课程,同时公共基础课程也要进行全面的学习。属性约简的结果可以提供给在校学生作为参考,为自己的就业目标有选择地进行学习与训练。受制于数据量的影响,得出的结论暂时不能泛化推广,但是对于在校学生还是有理论上的指导意义。对于目前存在的问题,将扩大现有数据集,并做进一步的研究。

参考文献:

[1] Pawlak Z. Rough sets[J]. International Journal of Computer & Information Sciences,1982,11(5):341-356.

[2] Pawlak Z. Rough sets:theoretical aspects of reasoning about data[M]. Dordrecht:Kluwer Academic Publishers,1991.

[3] Pawlak Z,Slowinski R. Routh set approach to multi-attribute decision analysis,invited review[J]. European Journal of Operational Research,1994,72(3):443-459.

[4] 王国胤,姚一豫,于　洪. 粗糙集理论与应用研究综述[J]. 计算机学报,2009,32(7):1229-1246.

[5] 胡清华,于达仁,谢宗霞. 基于邻域粒化和粗糙集逼近的数值属性约简[J]. 软件学报,2008,19(3):640-649.

[6] 胡清华,赵　辉,于达仁. 基于邻域粗糙集的符号与数值属性快速约简算法[J]. 模式识别与人工智能,2008,21(6):732-738.

[7] 李　墨. 粗糙集属性约简算法研究及其在大学生就业系统中的应用[D]. 广州:华南理工大学,2014.

[8] 计文军,蒋　超,王艳华,等. 粗糙集在大学生就业问题中的应用[J]. 内江师范学院学报,2008,23:232-234.

[9] 李彩虹. 基于粗糙集理论的大学生创业影响因素研究[J]. 技术与创新管理,2016,37(1):110-113.

[10] 叶东毅,黄翠微,赵　斌. 粗糙集中属性约简的一个贪心算法[J]. 系统工程与电子技术,2000,22(9):63-65.

[11] 崔建国,宋博翰,董世良,等. 基于邻域粗糙集的航空发电机健康诊断方法[J]. 数据采集与处理,2012,27(1):80-84.

[12] Liu Yong,Huang Wenliang,Jiang Yunliang,et al. Quick attribute algorithm for neighborhood rough set model[J]. Information Science,2014,271(7):65-81.

[13] 娄　畅,刘遵仁,郭功振. 基于块集的邻域粗糙集的快速约简算法[J]. 计算机科学,2014,41(11A):337-339.

[14] 贺超波,陈启买. 高校课程相关性粗糙集分析模型及应用[J]. 计算机工程与应用,2011,47(27):233-235.

(上接第 187 页)

[8] Ng D W K,Lo E S,Schober R. Multi-objective resource allocation for secure communication in cognitive radio networks with wireless information and power transfer[J]. IEEE Transactions on Vehicular Technology,2016,65(5):3166-3184.

[9] Pei Y,Liang Y C,Teh K C,et al. Secure communication in multiantenna cognitive radio networks with imperfect channel state information[J]. IEEE Transactions on Signal Processing,2011,59(4):1683-1693.

[10] Wu W,Zhang X,Wang S,et al. Max-min fair wireless energy transfer for MIMO wiretap channels[J]. IET Communication,2016,10(7):739-744.

[11] Liu L,Zhang R,Chua K C. Secrecy wireless information and power transfer with MISO beamforming[J]. IEEE Transactions on Signal Processing,2014,62(7):1850-1863.

[12] Zhang R,Ho C K. MIMO broadcasting for simultaneous wireless information and power transfer[J]. IEEE Transactions on Wireless Communication,2013,12(5):1989-2001.

[13] 孟庆民,龚家乐,曾桂根,等. 5G 多天线系统中毫米波物理层安全设计[J]. 计算机技术与发展,2016,26(2):91-94.

[14] Wang J,Palomar D P. Worst-case robust MIMO transmission with imperfect channel knowledge[J]. IEEE Transactions on Signal Processing,2014,57(8):3086-3100.

[15] Huang J,Swindlehurst A L. Robust secure transmission in MISO channels based on worst-case optimization[J]. IEEE Transactions on Signal Processing,2012,60(4):1696-1707.