

大数据时代下关联规则兴趣度挖掘在就业分析中的应用

李佐军

(滇西科技师范学院信息工程学院, 云南 临沧 677000)

摘要:近年来,随着大数据技术的应用普及,大数据正在改变各行业的发展。在教育领域,大数据技术在解决毕业生就业“难”和教育数据“闲置”问题方面产生了巨大的作用。本文在大数据时代背景下,使用Visual FoxPro开发了关联规则兴趣度挖掘工具,并使用该工具对影响学生就业的各项事务进行挖掘分析,找出影响就业的原因,为及时调整就业指导提供依据。

关键词:大数据;关联规则;兴趣度;就业策略

中图分类号: TP311.1 **文献标识码:** A

The Application of Interestingness Mining of Association Rules in Employment Analysis in the Era of Big Data

LI Zuojun

(School of Information Science & Engineering, Dianxi Science and Technology Normal University, Lincang 677000, China)

Abstract: In recent years, with the popularity of big data technology, big data is changing the development of various industries. In the field of education, big data technology plays a significant role in solving the difficulties of employment and processing the idle educational data. In the context of big data era, this paper uses Visual FoxPro to develop a mining tool of association rule interestingness, and uses this tool to mine and analyze the various issues affecting students' employment, finding out the factors that affect employment and providing evidence for timely adjustment of employment guidance.

Keywords: big data; association rules; interestingness; employment strategy

1 引言(Introduction)

随着计算机技术和信息技术的发展,大数据逐渐被应用于各行各业发展中,使用大数据技术采集和分析用户活动信息数据,掌握用户行为动态和爱好,为用户提供便捷的服务。利用大数据技术对毕业生就业相关事务数据进行详细的统计分析,找出影响毕业生就业的因素,及时调整就业策略,为毕业生顺利就业创造条件,切实提高就业率和就业质量。

2008年9月,《Nature》首次发布“大数据”专栏,引发了学术界对大数据的关注。当前“大数据”已成网络热点,受到了各界高度关注,作为信息社会发展的新事物,大数据正在逐步被认识和应用,大数据及时、高效、多样、量大及应用广泛的特点得到了各领域一致认可。麦肯锡全球研究所指出:“大数据是超出传统数据库软件获取、存储等能力的数据集,并不是大于一定数量的数据。因为随着科学技术的发展,大数据标准也会增加,不同行业标准也会发生变化”。

随着高等学校的发展和入学人数的增加,毕业生人数从

2000年的107万增加到2018年的820万,毕业生就业“难”的问题已成为学校必须面对的严峻挑战。同时,学校使用的各种信息管理软件在日常教学、管理和就业工作中积累了大量的管理数据,这些数据没有被利用,处在“闲置”状态。为了解决毕业生就业“难”和教育数据“闲置”的问题,学校可使用大数据技术对学生的个人、学习、就业等数据进行分析研究,找出影响就业的因素,及时调整就业指导和教育策略,引导学生顺利就业。

如今,大数据应用工具普遍存在,许多大数据工具都是在数据挖掘技术的基础上开发的。目前对数据挖掘技术的讨论尚未过时,研究数据挖掘技术是为了更好地利用大数据奠定基础。关联规则兴趣度挖掘正好能将“闲置”教育数据用起来,为学校各项教育决策服务。

2 关联规则挖掘及兴趣度(Association rule mining and interestingness)

2.1 关联规则挖掘

R.A Grawal等人在1993年的SIGMOD会议上提出了关联

规则挖掘,并将其定义为:设 $I=\{i_1, i_2, \dots, i_n\}$ 是项集, D 是事务集, T 是 I 上子集, $T \subseteq I$, TID是单个事务, 则关联规则表示为 $X \Rightarrow Y$, 其中 $X \subseteq I$, $Y \subseteq I$ 且 $X \cap Y = \emptyset$, X 叫做规则的条件, Y 叫做规则的结果^[1]。

为了提高关联规则挖掘的准确性, 引入了计算式为(1)和(2)的两个参数, 它们分别叫做支持度(S)和置信度(C)。

$$S(X \Rightarrow Y) = P(X \cup Y) = \frac{\text{Count}(X \cup Y)}{N} \times 100\% \quad (1)$$

$$C(X \Rightarrow Y) = P(Y|X) = \frac{S(X \Rightarrow Y)}{S(X)} = \frac{\text{Count}(X \cup Y)}{\text{Count}(X)} \times 100\% \quad (2)$$

其中, D 表示事务数据库, N 表示事务数的总和, 而 $\text{Count}(X)$ 表示事务 X 的发生次数, $\text{Count}(X \cup Y)$ 表示事务 X 和 Y 同时发生的次数。支持度 S 是事务 X 发生的频度, 置信度 C 是事务 X 发生时事务 Y 发生的频度。

2.2 兴趣度

在关联规则挖掘中, 仅使用支持 S 和置信度 C 的两个参数来测量规则的价值, 会导致部分强关联规则无使用价值。当最小支持阈值和最小置信度阈值太小时, 会生成冲突规则; 当太大时, 会遗漏有意义的规则^[2]。因此, 提出了一种新的度量关联度规则挖掘参数——兴趣度, 用来弥补支持度与置信度的缺陷, 提高数据挖掘结果的价值。

目前, 现成的关联规则挖掘兴趣度模型很多, 有概率兴趣度模型、Symth函数兴趣度模型、Gini指标兴趣度模型、Piantesky-Shapiro兴趣度模型、基于差异思想的兴趣度模型等^[3]。本文将使用一种由文献[4]提出的兴趣度模型, 具体描述如下:

设 $I=\{i_1, i_2, \dots, i_n\}$ 是项目集, D 是事务数据库, 关联规则 $X \Rightarrow Y$ 的兴趣度由(3)式计算, 兴趣度值越大关联规则挖掘结果越有趣越具有参考价值^[5]。

$$I(X \Rightarrow Y) = \frac{\frac{\text{Count}(X \cup Y)}{\text{Count}(X)} - \frac{\text{Count}(Y)}{N}}{\sqrt{\frac{\text{Count}(Y)}{N} \times \left(1 - \frac{\text{Count}(Y)}{N}\right)}} \quad (3)$$

3 关联规则兴趣度挖掘(Interestingness mining of association rules implementation)

3.1 数据准备

挖掘对象主要使用入学数据、学业成绩数据和就业数据, 通过筛选、合并等将这些数据组合成二维表, 并进行预处理。预处理后的就业事务数据表结构见表1所示。

表1 就业事务数据表结构

Tab.1 Structure of the employment transaction data table

学号	性别	专业	高考成绩	学业成绩	就业地区类型	单位性质
----	----	----	------	------	--------	------

3.2 算法实现

就业事务挖掘工具使用FoxPro作为后台数据库管理工具, 以Visual FoxPro 6.0作为系统开发平台。

3.2.1 数据库实现

就业事务表包括学号、性别、专业、高考成绩、学业成绩、就业地区类型、单位性质等七个属性, 其结构见表2

所示。

表2 就业事务表数据字典

Tab.2 Employment transaction table data dictionary

序号	字段名	类型	宽度	备注
1	学号	字符型	8	
2	性别	字符型	2	
3	专业	字符型	18	
4	高考成绩	字符型	8	总分
5	学分成绩	字符型	8	各学科平均值
6	就业地区类型	字符型	14	生源就业或非生源就业
7	单位性质	字符型	18	国有企业、私有企业、事业单位等

3.2.2 关联规则挖掘的实现

就业事务挖掘工具的操作对象是就业事务数据, 以文本的形式产生关联规则, 格式为: 属性A……→属性B…兴趣度: I。

关联规则挖掘算法由编程工具Visual FoxPro实现, 其算法流程图见图1所示。

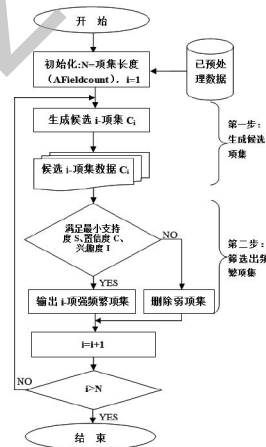


图1 关联规则兴趣度挖掘算法流程图

Fig.1 Association rule interestingness mining algorithm flow chart

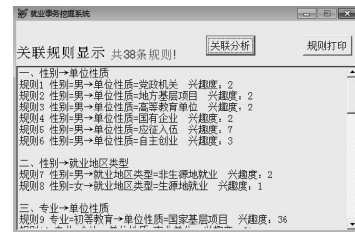


图2 就业事务挖掘结果显示

Fig.2 Employment transaction mining results display

4 关联规则兴趣度挖掘结果分析(Analysis of association rules interestingness mining results)

4.1 挖掘结果显示

根据预先设定的兴趣度阈值, 显示满足条件的规则, 其挖掘界面如图2所示。详细规则显示如下:

规则1 性别=男→单位性质=党政机关 兴趣度: 2

规则2 性别=男→单位性质=地方基层项目 兴趣度: 2

- 规则3 性别=男→单位性质=高等教育单位 兴趣度: 2
规则4 性别=男→单位性质=国有企业 兴趣度: 2
规则5 性别=男→单位性质=应征入伍 兴趣度: 7
规则6 性别=男→单位性质=自主创业 兴趣度: 3
规则7 性别=男→就业地区类型=非生源地就业 兴趣度: 2
规则8 性别=女→就业地区类型=生源地就业 兴趣度: 1
规则9 专业=初等教育→单位性质=国家基层项目 兴趣度: 36
规则10 专业=会计→单位性质=事业单位 兴趣度: 31
规则11 专业=计算机多媒体技术→单位性质=自主创业 兴趣度: 56
规则12 专业=计算机应用技术→单位性质=医疗卫生单位 兴趣度: 30
规则13 专业=历史教育→单位性质=应征入伍 兴趣度: 368
规则14 专业=美术教育→单位性质=升学 兴趣度: 33
规则15 专业=社区管理与服务→单位性质=事业单位 兴趣度: 37
规则16 专业=生物教育→单位性质=升学 兴趣度: 81
规则17 专业=体育教育→单位性质=地方基层项目 兴趣度: 32
规则18 专业=体育教育→单位性质=自主创业 兴趣度: 64
规则19 专业=物理教育→单位性质=国家基层项目 兴趣度: 30
规则20 专业=音乐教育→单位性质=升学 兴趣度: 39
规则21 专业=会计电算化→就业地区类型=生源地就业 兴趣度: 23
规则22 专业=历史教育→就业地区类型=生源地就业 兴趣度: 35
规则23 专业=学前教育→就业地区类型=生源地就业 兴趣度: 20
规则24 高考成绩=一般→单位性质=党政机关 兴趣度: 1
规则25 高考成绩=一般→单位性质=地方基层项目 兴趣度: 1
规则26 高考成绩=一般→单位性质=国家基层项目 兴趣度: 4
规则27 高考成绩=一般→单位性质=应征入伍 兴趣度: 1
规则28 高考成绩=优良→单位性质=高等教育单位 兴趣度: 1
规则29 高考成绩=优良→单位性质=国有企业 兴趣度: 4
规则30 高考成绩=优良→单位性质=私有企业 兴趣度: 1
规则31 高考成绩=一般→就业地区类型=生源地就业 兴趣度: 2
规则32 高考成绩=优良→就业地区类型=非生源地就业 兴趣度: 2
规则33 学业成绩=一般→单位性质=党政机关 兴趣度: 2
规则34 学业成绩=一般→单位性质=地方基层项目 兴趣度: 2
规则35 学业成绩=一般→单位性质=自主创业 兴趣度: 3
规则36 学业成绩=优良→单位性质=升学 兴趣度: 2
规则37 学业成绩=一般→就业地区类型=生源地就业 兴趣度: 1
规则38 学业成绩=优良→就业地区类型=非生源地就业 兴趣度: 1

4.2 挖掘结果分析

关联规则兴趣度挖掘主要任务是找出就业事务中各数据项间的关系,更好地服务于学校就业指导工作。对38条挖掘结果作进一步分析得出如下结论:

第一,规则1—6只显示了男生的就业领域,而没有显示女生的,说明男生就业范围比女生大。因此,学校相关就业管理部门要加大对女生的就业范围宣传和就业考试辅导,扩大女生就业面。

第二,从规则7—8可看出,男生主要在非生源地就业,而女生则主要在生源地就业。说明男生就业区域更广,而女生则正好相反,主要在生源地(家乡)就业。因此,学校要针对性的对女生进行就业思想观念教育,鼓励其到非生源地就业,扩大就业区域范围。

第三,从规则9—20可看出,各专业毕业生就业主要分布在升学、国家或地方项目、事业单位、自主创业等方面。因

此,学校要加强就业宣传,引进私有企业等单位到校招聘毕业生。

第四,从规则21—23可看出,显示了三个专业的就业地区类型都为“生源地就业”。由此可推断出,各专业毕业生主要就业地区为生源地。所以,学校要通过教育改变学生的就业观念,鼓励到非生源地就业,扩大就业区域范围。

第五,从规则24—30可看出,高考成绩“一般”和“优秀”的学生在各种企业单位基本都有就业。所以,高考分数高低对于学生就业的单位没有太大影响。

第六,从规则31、32可看出,高考成绩“一般”的学生主要回生源地就业,而高考成绩“优秀”的学生主要在非生源地就业。因此,学校要引导高考成绩“一般”的学生去非生源地就业,引导高考成绩“优良”的学生回生源地就业,扩大毕业生就业区域范围。

第七,从规则33—36可看出,在校期间学习成绩“一般”的学生就业领域为党政机关、地方基层项目、自主创业等,在校期间学习成绩“优良”的学生主要出路为升学。由此可知,在校期间学习成绩“一般”的学生就业领域更广,而学习成绩“优良”的学生更适合考试升学进修。因此,学校要对学习成绩一般的学生加强考试辅导,引导其参加升学考试;对学习成绩较好的学生除了继续进行升学考试辅导外,还要引导其向其他领域就业,促进各领域均衡就业。

第八,从规则37、38可看出,在校期间学习成绩“一般”的学生主要回生源地就业,而在在校期间学习成绩“优良”的学生主要去非生源地就业。由此可知,在校期间学习成绩好的学生更有“闯劲”,敢于挑战陌生环境。由此,学校要对学习成绩一般的学生开展就业观念教育,引导他们去非生源地就业。

5 结论(Conclusion)

本文在大数据时代背景下对关联规则兴趣度挖掘在毕业生就业分析中的应用进行了研究,把兴趣度引入了数据分析中,提高了数据挖掘结果的使用价值。使用Visual FoxPro开发的数据挖掘工具对学生就业各项事务数据进行关联规则兴趣度挖掘分析,找到各事务间存在的联系,对学校开展就业管理工作具有重要意义。同时为学校利用大数据技术分析就业数据增添了新的思路,为学校就业管理者开展工作提供新的依据。

参考文献(References)

- [1] Jiawei Han, Micheline Kamber[加]. 范明, 孟小峰, 译. 数据挖掘概念及技术[M]. 北京: 机械工业出版社, 2008: 147-154.
- [2] 董辉. 基于兴趣度的高职课程关联规则挖掘[J]. 吉首大学学报: 自然科学版, 2012, 33(3): 41-46.
- [3] 吴杰. 基于兴趣度的关联规则挖掘[D]. 哈尔滨工业大学, 2009: 41-42.
- [4] 李永立, 吴冲, 王崑声. 一种新的关联规则兴趣度度量方法[J]. 情报科学, 2011, 30(5): 503-507.
- [5] 李佐军. 基于关联规则兴趣度的课程设置研究[J]. 大理学院学报, 2014, 6: 20-23.

作者简介:

李佐军(1980—), 男, 硕士, 副教授. 研究领域: 数据挖掘, 大数据应用技术。