

基于 ID3 算法对农民工城市融入影响因素分析

陈韬宇¹, 安海燕¹, 陈 杰²

(1.贵州大学经济学院, 贵州 贵阳 550025;

2.贵州师范大学国际教育学院, 贵州 贵阳 550001)

✉ chenty713@126.com; 1240133318@qq.com; 870534478@qq.com



摘 要:在农民工逐步成为城市主体流动人口的过程中,其城市融入问题与社会的和谐发展息息相关。针对该问题,运用迭代二叉树 3 代(Iterative Dichotomiser 3, ID3)算法对影响农民工城市融入的影响因素进行探究,将影响因素分为经济、社交、社会福利和心理 4 个方面,采取与经典计量经济学不同的 ID3 算法对调研数据进行分析。结果显示,经济因素对农民工城市融入的影响最高,信息增益的结果达到 0.192,社会福利因素的影响次之,心理和社交因素对于城市融入的影响程度比较低。在此研究结果的基础上对农民工城市融入问题提出些许建议。

关键词:城市融入; ID3 算法; 农民工

中图分类号: TP312 **文献标志码:** A

Analysis of the Factors Influencing the Integration of Migrant Workers into Cities Based on ID3 Algorithm

CHEN Taoyu¹, AN Haiyan¹, CHEN Jie²

(1.School of Economics, Guizhou University, Guiyang 550025, China;

2.School of International Education, Guizhou Normal University, Guiyang 550001, China)

✉ chenty713@126.com; 1240133318@qq.com; 870534478@qq.com

Abstract: In the process of migrant workers gradually becoming the main mobile population in the city, their urban integration is closely related to the harmonious development of society. To solve the problem, this paper proposes to use Iterative Dichotomiser 3 (ID3) algorithm to explore the influencing factors that affect the integration of migrant workers into cities. The influencing factors are divided into 4 aspects: economic, social, social welfare and psychological. The ID3 algorithm, which is different from classical econometrics, is used to analyze the survey data. The results show that economic factors have the highest impact on urban integration of migrant workers, with an information gain of 0.192, followed by social welfare factors, and psychological and social factors have a relatively low impact on urban integration. Based on the research results, some suggestions are proposed for migrant workers' urban integration.

Key words: urban integration; ID3 algorithm; migrant workers

0 引言(Introduction)

近年来,在我国经济快速发展下,产生了特有的农民工群体。城镇化发展至今,农村劳动力转移的同时,也导致了各式各样的问题^[1-3]。从经济学的视角来看,不论是生活的本质——衣食住行,还是社会保障、教育等都会影响农民工在城

市的融入感^[4-6]。如今,新生代农民工逐渐成为老一代农民工的接班人,而不同的成长环境和经历,造成其在融入城市的过程中又产生了新的问题^[7-9]。因此,有必要对影响农民工城市融入的因素进行进一步探究。本文采用信息学的 ID3 算法分析经济、社交、社会福利和心理 4 个方面对农民工城市融入问

题的影响,并为解决农民工城市融入问题提出相应的建议^[10]。

本文以直观的数据结果分析影响农民工城市融入的因素。其中,决策树算法的典型应用在文献[11]至文献[13]中有提及,同时有许多学者对其进行了优化^[14-16]。本文以大数据挖掘为基础,对城市融入问题的研究进行方法上的创新^[17-18]。

有别于传统回归分析方式,采取数据挖掘算法 ID3 分析农民工的城市融入问题,对影响因子的重要性进行树形排列,得到直观的信息,为该问题的深入研究提供方向和参考。

1 ID3 算法模型 (ID3 algorithm model)

ID3 算法就是决策树模型算法,它是一种基本的分类回归方法,通常运用于大数据挖掘方面,与 ID3 算法类似的有 C4.5 算法 (ID3 算法的一种扩展算法) 和 CART 算法。C4.5 算法的优点在于产生的规则容易被理解,正确率比较高,操作实现的难度不高,但是它的缺点也同样明显,例如数据需要进行多次遍历和排序,效率很低。ID3 算法在剪枝和数据丢失值处理等方面区别于 C4.5 算法和 CART 算法。针对本文的问题,采取 ID3 算法更简洁、有效,能够准确地挖掘数据中隐藏的含义,体现影响农民工城市融入的主要因素和次要因素,也就是重要性的排序。

1.1 算法的介绍

ID3 算法是一种经典的决策树算法,常用于分析数据的重要性排序;它的原理是建立在奥卡姆剃刀原理的基础上,用较少的样本获取更多的隐藏信息。在信息理论中,期望信息越小,那么信息增益 (Gain) 就越大。ID3 算法的核心思想就是通过信息增益大小作为属性的选择理论依据进行分裂挖掘。

1.2 ID3 算法中数值的基本定义

(1) 信息熵。熵的概念是在物理学中被定义的,它是在物理学中用作测量一个热力学系统的无序程度,在信息学的理论中,熵是对不确定性信息的度量。香农 (Claude Shannon) 对物理学中的熵重新进行定义,出现了信息学中的信息熵,并将其定义为离散随机事件出现的概率。一个数据集或者系统当中的数据越有序,信息熵就越低;一个数据集或者系统当中的数据越无序,信息熵就越高。假如一个随机变量 D 的取值为 $D = \{d_1, d_2, d_3, d_4, \dots, d_n\}$, 每一种取到的概率分别是 $\{p_1, p_2, p_3, \dots, p_n\}$, 那么 D 的熵定义如下:

$$Entropy(D) = -\sum_{i=1}^n p_i \log_2 p_i \quad (1)$$

其中, $Entropy(D)$ 为数据总信息熵, p_i 为取到该项数据的概率。

一个变量的变化越多,那么它携带的信息量就越大。本文研究通过 ID3 算法对调研数据进行处理,分析调研数据的信息量并用信息熵表示。由熵的定义可知,熵只决定了 X 的分布规律,与 X 的取值大小没有关系。

(2) 期望信息也称为分类系统的熵或者条件熵,对于分类系统来说,特征 A 是拥有不同的类别 $\{A_1, A_2, \dots, A_n\}$ 变量,它的样本个数分别记为 $\{C_1, C_2, C_3, \dots, C_n\}$, 而每一个类别出现的概率分别是 $P(C_1), P(C_2), \dots, P(C_n)$, 其中 n 表示类别的总数量。所以,期望信息 (条件熵) 就可以用公式 (2) 表示:

$$H(D/A) = -\sum_{i=1}^n \frac{C_i}{C} P(C_{ij}) \log_2 P(C_{ij}) \quad (2)$$

其中, $H(D/A)$ 表示条件熵, $\frac{C_i}{C}$ 为特征 A 中的第 i 类别的数量

占特征 A 所有类别的概率, $P(C_{ij})$ 为特征 A 中的第 j 类别占样本集合 S 不同类别的概率。

(3) 信息增益是 ID3 算法中最核心的判别特征,也是区别重要性的依据。将信息增益定义为 $Gain$, 信息增益是数据集 D 分裂前与分裂后的差,表示由于特征的分裂引起数据集的不确定性程度降低。通过计算得到数据分裂的根据是由于当特征的取值较多时,以此特征划分更容易得到纯度更高的子集,表现为如图 1 所示的经济因素特征下还有特征可以分裂,也就是决策树根节点顺序由高到低表示重要程度的高低。

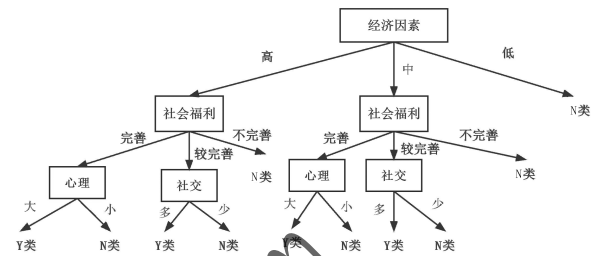


图1 决策树算法模型结果图

Fig. 1 Result graph of decision tree algorithm model

信息增益越高,意味着划分后的子节点纯度越高,对于分类的贡献越大。因此, ID3 算法须选择信息增益最大的节点作为父节点。表示信息增益的公式如下:

$$Gain_A(D) = E(D) - H(D/A) \quad (3)$$

其中, $Gain_A(D)$ 表示特征 A 中的信息增益, $E(D)$ 表示原始样本的信息熵, $H(D/A)$ 表示特征 A 的条件熵。

2 调研数据处理及分析 (Research data processing and analysis)

2.1 数据来源和数据的初步处理

数据来源于实地调研,有效的调研问卷总计 359 份,所有数据均来自调研对象农民工。在问卷中设计了包括经济 (收入)、社会福利 (教育和医疗水平)、社交 (亲友数量) 及心理 (受到歧视心理影响程度) 等方面的问题。将调查的实验数据进行汇总并归纳分析,将其分为 15 个大类,重复的数据归为一个类别,得到的归纳分析结果如表 1 所示。

表1 数据归纳分析表

Tab.1 Data summary and analysis table

类别	经济 (收入)	社会福利 (教育和医疗水平)	社交 (亲友数量)	心理 (受到歧视心理影响程度)	自我感觉融入
1	高(5)	较完善(3)	少(1)	小(1)	否(1)
2	高(5)	完善(5)	少(1)	小(1)	否(1)
3	高(5)	完善(4)	少(2)	大(3)	是(2)
4	高(5)	较完善(3)	多(4)	大(3)	否(1)
5	高(5)	较完善(3)	少(2)	小(1)	否(1)
6	中(4)	不完善(2)	多(4)	大(3)	否(1)
7	中(4)	完善(4)	多(3)	小(1)	否(1)
8	中(4)	完善(4)	少(1)	小(1)	否(1)
9	中(4)	较完善(3)	少(1)	小(1)	是(2)

续表

类别	经济 (收入)	社会福利(教育 和医疗水平)	社交 (亲友数量)	心理(受到歧视心 理影响程度)	自我感 觉融入
10	中(4)	完善(4)	少(2)	小(1)	是(2)
11	低(3)	完善(5)	少(2)	大(3)	是(2)
12	低(3)	较完善(3)	少(2)	大(3)	是(2)
13	低(3)	较完善(3)	多(5)	小(2)	是(2)
14	低(3)	完善(5)	多(4)	小(2)	是(2)
15	低(3)	不完善(2)	少(2)	小(1)	否(1)

本文研究将影响城市融入的影响因素分为经济(收入)、社会福利(教育和医疗水平)、社交(亲友数量)、心理(受到歧视心理影响程度)4个方面。问卷调查中将收入变化和生活成本变化程度由弱到强分为5个。但是,由于问卷中的问题过于细化,所以将5个程度减少为3个程度,分别是低、中、高,并且由于在调查问卷中调研对象选择的程度在3~5的范围,所以可以将程度1~2去除,最后的结果为低(3)、中(4)、高(5),分别代表了调研对象对经济部分问题的感知和反馈程度。社会福利、社交和心理三个影响因素进行同样的量化分类,具体如表1所示。

2.2 数据计算处理过程

2.2.1 计算总体熵

经归纳分析后的问卷结果见表1。表1中有15个例子,共有两个大类,即城市融入的“是”与“否”,同时拥有4个特征,分别经济、社会福利、社交和心理。感到融入“是”的比例为7/15,感到融入“否”的比例为8/15。

计算总体熵,假设 D 为表1所有的数据集合,经计算可得:

$$Entropy(D) = -\sum_{i=1}^n p_i \log_2 p_i = -\frac{7}{15} \log_2 \frac{7}{15} - \frac{8}{15} \log_2 \frac{8}{15}$$

这里将 $Entropy(D)$ 记为缩写 $E(D)$,所以 $E(D) \approx 0.997$ (保留三位小数,下同)。

2.2.2 计算特征下的条件熵

4个特征即 C_1 为经济、 C_2 为社会福利、 C_3 为社交、 C_4 为心理,将两个大类记为 B_1 与 B_2 ,分别用 A_1 、 A_2 和 A_3 代表经济因素影响的高、中、低三个程度。

(1)以经济为特征计算条件熵(如表2所示)。

表2 经济特征条件熵表

Tab.2 Entropy table of economic characteristic conditions

经济因素 比例 (C_1 为经济)	经济因素占总 体因素的比例 PC_1	经济因素影响下 感到融入社会的 比例 $P(B_1/C_1A_i)$	经济因素影响下 未感到融入社会的 比例 $P(B_2/C_1A_i)$
C_1A_1	5/15	1/5	4/5
C_1A_2	5/15	2/5	3/5
C_1A_3	5/15	4/5	1/5

根据经济为特征的熵:

$$H(D/C_1) = PC_1 \times H(D/C_1A_1) + PC_1 \times H(D/C_1A_2) + PC_1 \times H(D/C_1A_3) \quad (4)$$

根据公式(2)可得样本特征为经济的高、中、低的熵:

$$H(D/C_1A_1) = -\frac{1}{5} \log_2 \frac{1}{5} - \frac{4}{5} \log_2 \frac{4}{5} \approx 0.722$$

$$H(D/C_1A_2) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \approx 0.971$$

$$H(D/C_1A_3) = -\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5} \approx 0.722$$

以经济为特征的熵:

$$H(D/C_1) = 0.722 \times \frac{1}{3} + 0.971 \times \frac{1}{3} + 0.722 \times \frac{1}{3} \approx 0.805$$

(2)以社会福利为特征计算条件熵(如表3所示)。

表3 社会福利条件熵表

Tab.3 Entropy table of social welfare conditions

社会福利因素 比例 (C_2 为社会福利)	社会福利因素 占总体因素的 比例 PC_2	社会福利因素 影响下感到融入 社会的比例 $P(B_1/C_2A_i)$	社会福利因素 影响下未感到融 入社会的比例 $P(B_2/C_2A_i)$
C_2A_1	7/15	4/7	3/7
C_2A_2	6/15	3/6	3/6
C_2A_3	2/15	0	1

根据社会福利为特征的条件熵:

$$H(D/C_2) = PC_2 \times H(D/C_2A_1) + PC_2 \times H(D/C_2A_2) + PC_2 \times H(D/C_2A_3) \quad (5)$$

$$H(D/C_2A_1) = -\frac{4}{7} \log_2 \frac{4}{7} - \frac{3}{7} \log_2 \frac{3}{7} \approx 0.985$$

$$H(D/C_2A_2) = -2 \times \frac{1}{2} \log_2 \frac{1}{2} = 1$$

$$H(D/C_2A_3) = -0 \log_2 0 - \log_2 1 = 0$$

以社会福利为特征的熵:

$$H(D/C_2) = 0.985 \times \frac{7}{15} + 1 \times \frac{6}{15} \approx 0.860$$

(3)以社交为特征计算条件熵(如表4所示)。

表4 社交特征条件熵表

Tab.4 Entropy table of social feature conditions

社交因素 比例 (C_3 为社交)	社交因素占 总体因素的 比例 PC_3	社交因素影响 下感到融入社会 的比例 $P(B_1/C_3A_i)$	社交因素影响 下未感到融入 社会的比例 $P(B_2/C_3A_i)$
C_3A_1	5/15	2/5	3/5
C_3A_3	10/15	5/10	5/10

根据社交为特征的条件熵:

$$H(D/C_3) = PC_3 \times H(D/C_3A_1) + PC_3 \times H(D/C_3A_3) \quad (6)$$

$$H(D/C_3A_1) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \approx 0.971$$

$$H(D/C_3A_3) = -2 \times \frac{1}{2} \log_2 \frac{1}{2} = 1$$

以社交为特征的熵:

$$H(D/C_3) = 0.971 \times \frac{5}{15} + 1 \times \frac{10}{15} \approx 0.990$$

(4)以心理为特征计算条件熵(如表5所示)。

表5 心理特征条件熵表

Tab.5 Entropy table of psychological feature conditions

心理因素 比例 (C_4 为心理)	心理因素占 总体因素的 比例 PC_4	心理因素影响下 感到融入社会的 比例 $P(B_1/C_4A_i)$	心理因素影响下 未感到融入社会的 比例 $P(B_2/C_4A_i)$
C_4A_1	5/15	3/5	2/5
C_4A_3	10/15	4/10	6/10

根据心理为特征的条件熵:

$$H(D/C_4)=PC_4 \times H(D/C_4A_1)+PC_4 \times H(D/C_4A_3) \quad (7)$$

$$H(D/C_4A_1)=-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5}\approx 0.971$$

$$H(D/C_4A_3)=-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}\approx 0.971$$

以心理为特征的熵:

$$H(D/C_4)=0.971 \times \frac{5}{15}+0.971 \times \frac{10}{15}=0.971$$

2.2.3 计算信息增益

信息增益通过以下公式计算,并选取出最大的信息增益作为树的根节点,也就是最重要因素的决出。

$$Gain(D/C_i)=E(D)-H(D/C_i) \quad (8)$$

$$Gain(D/C_1)=E(D)-H(D/C_1)=0.997-0.805=0.192$$

$$Gain(D/C_2)=E(D)-H(D/C_2)=0.997-0.860=0.137$$

$$Gain(D/C_3)=E(D)-H(D/C_3)=0.997-0.990=0.007$$

$$Gain(D/C_4)=E(D)-H(D/C_4)=0.997-0.971=0.026$$

选取具有最大增益的特征作为树的根节点,也就是经济因素 C_1 作为根节点。

(1)接下来进行数据集划分,经济因素影响程度高(C_1A_1)对应表1的数据集为 $X1=\{1,2,3,4,5\}$;经济因素影响程度中(C_1A_2)对应表1的数据集为 $X2=\{6,7,8,9,10\}$;经济因素影响程度低(C_1A_3)对应表1的数据集为 $X3=\{11,12,13,14,15\}$ 。

(2)进行数据集的划分后,对每个数据集进行决策树算法,在数据集的子集中求出各个子集信息增益,从而得到 $X1$ 中 C_2 特征最为显著,也就是信息增益最大,将其设置为分枝根节点。同理,在 $X2$ 中的信息增益最大的也是 C_2 ,对 C_2 再进行决策树算法,同时对其数据集划分得到 C_2 特征中高的信息增益为特征 C_4 , C_2 特征中的信息增益较高的为 C_3 ,所以得到信息增益在各个特征下的信息增益排序,各个特征下低的部分皆为否类,并画出决策树的图(如图1所示)。

3 结论(Conclusion)

根据实验结果和数据决策树分析图可知,经济因素是第一影响因素,社会福利因素是次要影响因素,这反映了农民工对社会福利的关心程度较高,从侧面反映了农民工只有在强烈感受到社会保障(如医疗和教育)的福利之后,才会降低城市无法融入感,甚至认可自身融入城市。对政府来说,完善社会福利体系,医疗、教育、保险都是很重要的方面。同时,对于社会福利的微观方面,应该尽力帮助农民工对享受到保障措施的程序简单化,杜绝各种享受福利需要各种部门盖章的烦琐程序,让农民工感受到城市对其的包容。研究表明,心理和社交因素对于农民工城市融入的影响程度比较低。农民工是否觉得自己

受到歧视以及影响其城市融入的程度并不是那么在意,更多的是关心自己的经济收入和社会福利是否得到保障。政府需要加强农民工的心理健康建设,例如让农民工重视工作的长远规划,建立职业信心与生活信念等。

参考文献(References)

- [1] 陈博欧,张锦华. 社交能力与农民工工资性收入[J]. 财经研究,2021,47(11):124-138.
- [2] 刘静,张锦华. 城市异质影响下的农民工市民化程度:基于需求可识别双变量 Probit 和 HLM 模型的测度与分析[J]. 浙江社会科学,2021(10):63-71,80.
- [3] 郭凤鸣. 青年和中年农民工过度劳动差异的成因分析[J]. 中国青年研究,2021(10):103-111.
- [4] 周云波,黄云. 基本医疗保险制度能否改善农民工的相对不平等[J]. 财经科学,2021(10):84-97.
- [5] 冯喜良,高盼盼,罗荣波. 互联网使用对农民工性别工资收入差距的影响[J]. 人口与经济,2021(5):111-124.
- [6] 倪超军. 城市公共服务开放度与农民工流迁行为[J]. 产经评论,2021,12(5):131-146.
- [7] 陈诚,杨巧. 城市视角下农民工居留和落户意愿分异特征及影响因素研究[J]. 华中科技大学学报(社会科学版),2021,35(5):69-80,113.
- [8] 吴丽娟,罗淳. 心理健康与进城农民工的多维贫困[J]. 华南农业大学学报(社会科学版),2021,20(1):84-95.
- [9] 苗海民,张顺莉,朱俊峰. 农民工家属选择性迁移对土地流转的影响:基于中国流动人口动态监测调查数据的经验分析[J]. 中国农村经济,2021(8):24-42.
- [10] 王子京,刘毓. 决策树 ID3 新属性选择方法[J]. 现代电子技术,2018,41(23):9-12.
- [11] 王子京,刘毓. 一种基于决策树 ID3 的改进算法[J]. 现代电子技术,2018,41(15):39-42.
- [12] 蒲天添. 基于决策树的工程项目管理优化研究[J]. 现代电子技术,2018,41(1):169-172.
- [13] 周云波,黄云. 基本医疗保险制度能否改善农民工的相对不平等[J]. 财经科学,2021(10):84-97.
- [14] 王世东,刘毅,王新闯,等. 基于改进决策树模型的矿区土地复垦适宜性评价[J]. 中国水土保持科学,2016,14(6):35-43.
- [15] 潘大胜,屈迟文. 一种改进 ID3 型决策树挖掘算法[J]. 华侨大学学报(自然科学版),2016,37(1):71-73.
- [16] 董跃华,刘力. 基于相关系数的决策树优化算法[J]. 计算机工程与科学,2015,37(9):1783-1793.
- [17] 石雁冰,邵勇,李树法. 基于 ID3 算法的代谢综合征主要因素分析[J]. 中国老年学杂志,2014,34(1):180-182.
- [18] DE RAUW A,ROZELLE S. Migration and household investment in rural China[J]. China Economic Review, 2008,19(2):320-335.

作者简介:

陈韬宇(1998-),男,硕士生。研究领域:农村区域发展。本文通信作者。

安海燕(1979-),女,博士,教授。研究领域:农业经济。

陈杰(1981-),女,硕士,讲师。研究领域:农业经济。