计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(8):190−194 [doi: 10.15888/j.cnki.csa.005906]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

应用自然邻居分类算法的大学生就业预测模型①

朱庆生, 高　璇

(重庆大学 计算机学院, 重庆 400044)

摘　要: 针对因大学生对薪酬预期过高而导致就业难的问题, 利用基于自然邻居的分类算法对近三年信息类专业毕业生的就业数据进行分析, 建立了大学生就业薪酬预测模型. 首先采用因子分析方法提取出决定大学生就业薪酬级别的潜在因子并作为模型输入变量, 进而应用基于自然邻居的分类算法对就业薪酬进行分类预测. 其中, 自然邻分类算法成功避免了 KNN 算法中存在的 K 值选取难题, 且每个节点的邻居数目会根据数据集的分布状况自适应获取. 实验结果表明, 该模型的预测精度高达 80.16%, 对于帮助大学生建立合理就业预期、提高就业能力等方面具有一定指导意义.

关键词: 数据挖掘; 自然邻居; 分类; 因子分析; 就业预测

引用格式: 朱庆生,高璇.应用自然邻居分类算法的大学生就业预测模型.计算机系统应用,2017,26(8):190–194. http://www.c-s-a.org.cn/1003-3254/5906.html

Model of College Students' Emolument Prediction Based on the Classification Algorithm with Natural Neighbor

ZHU Qing-Sheng, GAO Xuan

(School of Computer Science and Technology, ChongQing University, Chongqing 400044, China)

Abstract: To solve the problem of hard employment of graduates who expect for the impractical emolument, the paper builds a model for emolument prediction. On the basis of an classification algorithm with natural neighbor(NaN)，it analyzes the employment data of graduates majoring in Information Engineering in past three years. The paper uses factor analysis method to fetch the latency of employment emolument level determinants. Classification predicts the emolument by applying the latency as a variable based on the classification algorithm. This algorithm avoids the difficulty of parameter selection in K-nearest neighbor(KNN). The neighbors of each node can also be acquired as the topography of data set. According to the experiments, the prediction accuracy is 80.16%. The paper can guide graduates to build a reasonable emolument prediction or improve employment.

Key words: data mining; natural neighbor; classification; factor analysis; emolument prediction

随着我国高等教育由精英教育转变为大众教育, 高校毕业生就业形势日趋严峻. 大学生就业日益困难的一个重要原因就在于就业预期偏高, 且主要表现在对于薪酬的预期过高. 因而对毕业生的就业能力进行准确评估, 制定合理的薪酬预期就显得十分重要. 但由于薪酬预测涉及经济学、管理学、人工智能等多学科领域, 影响因素众多, 实现较为困难, 因此这项研究在成为广大研究人员关注热点的同时, 也成为一项难点研究课题[1-3].

文献[4]系统地分析大学生就业能力结构, 提出影响就业质量的三个主要因子为“就业人格”、“准职业形象”以及“社会兼容度”, 但没有对指标进一步分析和

① 收稿时间: 2016-12-07; 采用时间: 2017-01-04

预测, 无法在实际生活中起到直接指导大学生择业、就业的作用. 文献[5]采用多项式回归和多元线性回归两种不同的算法实现了对高校就业率的评估和预测, 但预测结果缺乏因果解释且无法对大学生个体的就业质量进行预测. 文献[6]提出采用一种基于信息增益比的决策树构造算法建立大学生就业预测模型, 但该模型无法处理包含高分支属性和噪声数据的真实数据集, 对训练拟合样本有着较大的依赖, 模型泛化能力欠佳.

自然邻居[7,8]是一种无尺度的最近邻居概念, 通过算法 1 自适应地计算出需要的参数, 具有无需人为输入参数、计算精准率高等优势.

针对上述问题以及自然邻居自适应的优势, 本文提出一种基于自然邻居分类算法的大学生就业预测模型. 本文通过高校就业信息管理系统获取近三年信息类专业毕业生就业情况, 创建了当前最具时效性的就业薪酬预测数据集, 进而结合因子分析法和基于自然邻居的分类算法建立模型预测毕业生的就业薪酬等级, 给出薪酬预测数学模型及实验结果, 并与目前广泛使用的经典分类算法进行性能分析对比, 探讨基于自然邻居的分类算法在大学生薪酬预测中的应用优势, 以期达到帮助大学生建立合理的就业预期, 提高就业率的最终目的.

1 基于因子分析法的毕业生描述模型建立

由于大学毕业生个体情况差异性很大, 很难对于单个个体薪酬的具体数值进行准确预测, 因此需要建立适当的毕业生薪酬等级, 能够基本符合毕业生实际薪酬差异水平. 本文根据个体样本薪酬与样本集平均薪酬的比例差异作为标准划分薪酬等级, 划分方式如表 1 所示, 其中 S_t 表示单个样本的薪酬, S_d 表示所有样本数据的平均薪酬.

表 1 薪酬等级表

评价指标	评价标准	评价级别
劳动报酬	$S_t \geq 1.5S_d$	A
	$1.5S_d > S_t \geq 1.2S_d$	B
	$1.2S_d > S_t \geq 0.8S_d$	C
	$0.8S_d > S_t \geq 0.5S_d$	D
	$0.5S_d > S_t$	E

在建立了薪酬等级之后, 本文在分析影响薪酬因素的基础上, 将样本特征数值化, 并利用因子分析方法简化影响因子, 建立能够描述样本主要特性的特征向量模型, 为后续毕业生薪酬等级分类奠定样本数据基础.

因子分析[9-11]的基本思想是将观测变量进行分类, 将相关性较高, 即联系比较紧密的分在同一类中, 而不同类变量之间的相关性则较低, 那么每一类变量实际上就代表了一个基本结构, 即公共因子. 这样, 就能相对容易地以较少的几个因子反映原资料的大部分信息, 从而达到浓缩数据, 以小见大, 抓住问题本质和核心的目的.

通过参考现有文献[12]形成的概念界定, 本文将 18 个可观测变量作为决定大学生就业薪酬级别的影响因子, 有 GPA、英语成绩、课程设计、兼职经历、学生干部、竞赛获奖、奖学金情况、实践项目、荣誉称号、校园活动、抗压能力、社交情况、政治面貌、参赛情况、实习经历、家庭状况、毕业设计成绩、表达能力. 同时为了使样本模型能够更加客观的对样本特征进行描述, 将不是数值型的观测变量转换成数值型数据, 并对所有观测变量数据进行归一化操作, 将全部数据归一到[0, 1]之间, 以消除数据量纲的差异, 归一化公式如下所示:

$$x_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

式中, x_i 表示归一化后的因素数据, $x_{\max}$ 表示因素变量最大值, $x_{\min}$ 表示因素变量最小值.

在对样本观测变量数值化的基础上, 利用因子分析法建立因子旋转成分矩阵表, 将 18 个影响薪酬水平的观测变量归纳出 4 个公共因子, 并以这 4 个公共因子对原始样本数据信息进行表征. 4 个因子分别包含以下变量指标.

因子 1: 包含 GPA 成绩、英语成绩、学术研究情况、奖学金获得情况, 体现了大学生在未来职业中所需的基本理论知识及快速学习的能力. 可以解释和命名为学生的“学习能力”.

因子 2: 包含课程设计成绩、项目经验、获奖经历、获奖级别, 体现了大学生在求职及未来工作过程中将所学技能应用于实践的能力. 可以解释和命名为学生的“实践能力”.

因子 3: 包含学生干部经历、干部级别、校园荣誉

称号获得情况、政治面貌, 体现了学生妥善处理组织内外关系的能力. 可以解释和命名为学生的“人际处理能力”.

因子 4: 包括兼职经历、实习经历、抗压能力, 体现了大学生在职场中快速适应社会角色转变以及严格执行岗位要求的能力. 可以解释和命名为“职业能力”.

将学习能力、实践能力、人际处理能力、职业能力 4 个维度作为探讨学生薪酬评级的主要影响因素. 毕业生样本的特征向量模型就可以表示为如下所示:

$$S_i = (Z_i, Y_i, J_i, L_i) \quad (2)$$

其中 S_i 表示第 i 个毕业生样本薪酬等级标签, Z_i, Y_i, J_i, L_i 分别表示第 i 个样本的学习能力、实践能力、人际处理能力、职业能力. 在进行薪酬等级预测分类之前, 需要将所有样本进行预处理, 将观测变量都转化为特征向量.

2 基于自然邻居分类的薪酬等级预测

在建立毕业生薪酬水平等级和毕业生样本特征向量的基础上, 本文采用基于自然邻居的分类算法对毕业生测试样本进行薪酬等级划定分类, 按照其特征向量将其映射到某一等级薪酬水平上去, 从而帮助毕业生确定合理的职业规划及准确的薪酬预期.

2.1 自然邻居搜索算法数

自然邻居(Natural Neighbor, NaN)作为一种特殊的 k-最近邻居(k Nearest Neighbor, KNN[13]), 它能够淡化 k 值的概念降, 低 KNN 算法分类性能与结果对于参数 K 选择的依赖. 即每个点邻居个数的计算过程不需要任何的参数, 可以由算法自适应计算产生. 自然邻居选择策略的核心思想是: 分布稠密的数据对象拥有较多的自然邻居, 而分布稀疏的数据对象则具有较少的自然邻居, 数据对象的疏密程度由覆盖该点的最近邻域数来确定.

算法 1 描述了自然邻居搜索算法的详细过程.

算法 1: 自然邻居搜索算法

输入: 数据集 X

输出: NaN_Edge, $NaN\_Num(x_i)$

算法步骤:

1 : $r = 1, flag = 0, NaN\_Edge = \emptyset$
2 : $Create\ a\ k-d\ tree\ T\ from\ data\ set\ X$
3 : $\forall x_i \in X, NaN\_Num(x_i) = 0$
4 : $While\ \ flag == 0\ \ do$
5 : $\quad for\ all\ \ x_i \in X\ \ do$
6 : $\quad\quad knn_r(x_i) = findKNN(x_i, r, T)$
7 : $\quad\quad KNN_r(x_i) = KNN_r(x_i) \cup \{knn_r(x_i)\}$
8 : $\quad\quad if\ \ x_i \in KNN_r(knn_r(x_i))\ \&\&$
$\{knn_r(x_i), x_i\} \notin NaN\_Edge\ \ then$
9 : $\quad\quad\quad NaN\_Edge = NaN\_Edge \cup$
$\{knn_r(x_i), x_i\}$
10 : $\quad\quad\quad NaN\_Num(x_i) = NaN\_Num(x_i) + 1$
11 : $\quad\quad\quad NaN\_Num(knn_r(x_i)) =$
$NaN\_Num(knn_r(x_i)) + 1$
12 : $\quad\quad end\ \ if$
13 : $\quad\quad endfor$
14 : $\quad\quad cnt = count(NaN\_Num(x_i) == 0)$
15 : $\quad\quad rep = repeat(cnt)$
16 : $\quad\quad if\ \ all(NaN\_Num(x_i)) \neq 0 || rep \geq \sqrt{r - repeat}$
$then$
17 : $\quad\quad\quad flag = 1$
18 : $\quad\quad endif$
19 : $\quad\quad r = r - 1$
20 : $\quad end\ \ while$
21 : $\quad \lambda = r - 1$
22 : $\quad Return : \lambda, NaN\_Edge, NaN\_Num(x_i)$

其中, NaN_Num 为集合 X 中每个点的自然邻居数; NaN_Edge 为自然邻居边集, 每一条边包含 2 个点; $findKNN(x_i, r, T)$函数为获取点 i 的第 r 个最近邻居; 函数 $count(NaN_Num(x_i))$计算自然邻居数; 函数 $repeat(cnt)$ 获取变量 cnt 值的重复次数, 当满足第 16 行所示条件时, 计算终止.

2.2 训练样本权重的分配算法

对于不规则的数据集, 在计算自然邻居时, 有可能出现一个样本的自然邻域中包含多个类别的情况, 此时分类结果会受到很大影响. 因此我们通过基于自然邻居的训练集加权算法[14]来提高分类精度.

假设测试集$T = \{t_1, t_2, \cdots, t_n\}$, 训练集$D = \{d_1, d_2, , \cdots, d_n\}$对于$\forall d_i \in D$, 设 $NaN(d_i)$表示样本 d_i 的自然邻域.

对于$\forall d_j \in NaN(d_i)$, 如果 d_j 与 d_i 的类标记相同, 则 d_j 就是 d_i 的好的自然邻居. 那么对于 d_i, 好的自然邻居的个数表示为 $Gnb(d_i)$.

对于$\forall d_j \in NaN(d_i)$, 如果 d_j 与 d_i 的类标记不相同, 则 d_j 就是 d_i 的坏的自然邻居. 那么对于 d_i, 坏的自然邻居的个数表示为 $Bnb(d_i)$.

坏的自然邻居常常为其他的样本提供了错误的分

类信息, 为了降低它们对分类结果的负面影响, 本文对训练样本集使用了如公式 3、4 所示的权重模式.

$$h_B(d_i)=\frac{Bnb(d_i)-\mu_{Bnb}}{\sigma_{Bnb}} \tag{3}$$

$$w(d_i)=\exp(-h_B(d_i)) \tag{4}$$

其中表示 Bnb 的平均值, σ_{Bnb} 表示 Bnb 的标准差, $h_B(d_i)$表示训练样本 d_i 的标准坏值, $w(d_i)$表示训练样本 d_i 所赋予的权重值.

算法 2: 基于自然邻居的训练集加权算法

输入: 训练样本集 D

输出: 训练样本集的权重矩阵

算法步骤:

Step1: 使用自然邻居搜索算法得到数据集 D 中每个样本的自然邻域;

Step2: 遍历每个样本的自然邻居得到每个样本的 *Gnb* 和 *Bnb*;

Step3: 根据已计算出的 *Bnb*, 计算其平均值μ_{Bnb} 和标准差σ_{Bnb};

Step4: 根据公式 3 中的加权修正因子, 得到每个样本的权重值, 输出权重矩阵.

2.3 基于自然邻分类算法的薪酬等级预测

将学生薪酬级别表示为特征向量后, 两个样本之间的相似度采用欧几里得距离来表示, 如下式所示, 其中, $d_i \in D$, S 表示样本的维度.

$$sim(t_i,d_j)=\frac{\sum_{k=1}^{s} t_{ik}\times d_{jk}}{\sqrt{\left(\sum_{k=1}^{s} t^2{}_{ik}\right)\times\left(\sum_{k=1}^{s} \mathrm{d}^2{}_{jk}\right)}} \tag{5}$$

结合公式 5, 根据如下的公式 6、7 统计出各类的自然邻居与测试样本的总相似度:

$$\delta(d_j,c_k)=\begin{cases}1 & if\ d_j\in c_k\\ 0 & if\ d_j\notin c_k\end{cases} \tag{6}$$

$$score(t_i,c_k)=\sum_{d_j\in NNN(t_i)} w(d_j)*sim(t_i,d_j)\delta(d_j,c_k) \tag{7}$$

基于自然邻居的分类算法的具体步骤如算法 3 描述所示.

算法 3: 基于自然邻居的分类算法

输入: 训练样本集 D, 测试样本集 T

输出: 带有类标记的训练样本集

算法步骤:

1 : $forall t_i \in T do$
2 : $Z=t_i\cup D$
3 : $(\lambda, NaN(t_i), NaN\_Num(t_i))=NaN(Z)$
4 : $if NaN\_Num(t_i)=0$
5 : $NaN(t_i)=KNN(t_i,\lambda)$
6 : $endif$
7 : $labelNum(T)=\max(score(t_i,c_k))$
8 : $returnlabelNum(T)$
9 : $endfor$

3 实验结果与分析

为了验证本文方法的有效性和准确性, 在课题组采集整理的数据集上进行毕业生薪酬预测实验, 数据集中样本个数为 1580, 每个样本包含了 18 种观测变量, 在数据预处理阶段已经将其转化为归一化的特征向量. 并选用 KNN 算法和加权 KNN 算法[15]作为实验对比算法, 利用预测准确率为客观评价指标来进一步对算法性能进行比较. 预测准确率计算公式如式(8)所示:

$$\text{准确率}=\frac{\text{结果分类正确样本数}}{\text{测试样本数}} \tag{8}$$

为了增强对比实验的可靠性, 实验在数据集上使用十折交叉验证法对数据集进行测试. 十折交叉验证法是将数据集分成十份, 轮流将其中 9 份作为训练集, 1 份作为测试集进行实验, 每次实验都会得出相应的准确率, 并对十次交叉实验结果的准确率进行平均作为算法精度的估计, 十折交叉验证实验结果如表 2 所示.

表 2 三种算法预测性能比较

次数	KNN	hw-KNN	本文算法
1	0.7518	0.7554	0.7691
2	0.7404	0.7424	0.7502
3	0.7648	0.7722	0.7794
4	0.7645	0.7832	0.7888
5	0.7634	0.8078	0.8245
6	0.7601	0.7988	0.8054
7	0.7798	0.8013	0.8105
8	0.7821	0.8164	0.8267
9	0.7728	0.8098	0.8148
10	0.7865	0.8204	0.8467
平均值	0.76662	0.79077	0.80161

由表 2 实验结果可以看出, 基于自然邻居的分类算法在就业薪酬预测中的分类精确度比其他两种方法

都高, 同时由于 KNN 和加权 KNN 算法检测结果是在设置不同参数经过多次实验后取得的最好结果, 而基于自然邻居的分类算法却无需设置任何参数, 算法复杂度和分类方法简便程度上还有着明显优势. 因此, 本文算法性能在就业薪酬预测方面明显优于传统算法.

4 结语

本文深入研究了当前影响大学生就业的主要因素, 将基于自然邻的分类算法应用到毕业生就业薪酬预测分析中, 建立了毕业就业薪酬预测分析模型. 该模型通过因子分析提取出四个决定大学生就业薪酬的潜在因子: 学习能力、实践能力、人际处理能力、职业能力, 并应用自然邻分类算法实现了对 5 种薪酬级别的预测. 通过与传统 KNN 算法和加权 KNN 算法的实验对比, 表明本文算法能够实现对毕业生薪酬等级的准确预测, 为高校管理人员以及大学生等提供就业指导的重要参考依据.

参考文献

1 潘文庆. 就业价值观对大学生就业质量的影响研究. 广东社会科学, 2014, (4): 40–46.

2 卫铁林. 基于 AHP 的高校毕业生就业质量评价模型构建. 郑州航空工业管理学院学报, 2013, (3): 90–94.

3 宋林, 张丛. 劳动力市场分割下大学生低水平就业的困境解析. 西北大学学报(哲学社会科学版), 2012, 42(1): 117–122.

4 贾利军, 管静娟. 大学生就业能力结构研究. 教育发展研究, 2013, (13): 51–56.

5 张稳, 恰汗•合孜尔. 毕业生就业率预测及质量评估研究. 计算机工程与科学, 2009, 31(5): 141–143.

6 蔡丽艳, 马弘伟. 数据挖掘技术在高校就业预测分析中的应用. 微计算机信息, 2012, 28(8): 101–103.

7 Zou XL, Zhu QS, Jin YF. An adaptive neighborhood graph for LLE algorithm without free-parameter. International Journal of Computer Applications, 2011, 16(2): 20–23. [doi: 10.5120/ijca]

8 Zhu QS, Huang JL, Feng J, *et al*. A clustering algorithm based on natural nearest neighbor. Journal of Computational Information Systems, 2014, 10(13): 5473–5480.

9 Bollen KA. Structural equations with latent variables. New York, N.Y.: Wiley, 1989.

10 Chatterjee S. Structural equation modeling: A bayesian approach. Technometrics, 2008, 50(3): 411–412.

11 Corrado L. Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models. Journal of the American Statistical Association, 2005, 100(470): 710–711. [doi: 10.1198/jasa.2005.s25]

12 王世通, 谢爱国. 基于因子分析的大学生就业能力影响因素研究. 当代经济, 2013, (21): 140–142.

13 Tomašev N, Radovanović M, Mladenić D, *et al*. A probabilistic approach to nearest-neighbor classification: Naive hubness bayesian kNN. Proc. 20th ACM International Conference on Information and Knowledge Management ACM. New York, NY, USA. 2011. 2173–2176.

14 Zhu QS, Zhang Y, Liu HJ. Classification algorithm based on natural nearest neighbor. Journal of Information and Computational Science, 2015, 12(2): 573–580. [doi: 10.12733/issn.1548-7741]

15 Radovanović M, Nanopoulos A, Ivanović M. Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research, 2010, 11(9): 2487–2531.