利用多元逻辑回归对 IT 行业薪酬变化的研究

高云龙

大连理工大学 辽宁省 大连市 116024
yunlong@mail.dlut.edu.cn

摘 要 基于调查数据研究软件行业薪资水平的影响因素,可帮助求职者预估薪资情况,也可为软件公司制定薪资方案提供理论依据。文章通过对某招聘网站进行数据爬取，获得近万条数据，对数据进行适当处理，把工作岗位、公司福利、公司类型、公司规模、学历和薪酬定性数据进行定量化。通过分析影响薪资的因素，之后建立多元线性回归模型刻画工作岗位、公司福利、公司类型、公司规模和学历对薪资的具体影响作用。

关键词： 大数据；IT 行业薪资影响因素分析；多元逻辑回归；薪资预测

中图法分类号 TP391.1

1 引言

在 Internet 飞速发展的今天，互联网已成为人们快速获取、发布和传递信息的重要渠道，它在政治、经济、生活等各个方面发挥着重要的作用，已成为政府、企事业单位信息化建设中的重要组成部分，从而倍受人们的重视。随着中国软件业的迅猛发展，IT 行业工作人员的薪资也“水涨船高”，这是吸引大量毕业生进入软件行业的主要动力。[1]本文调查了某招聘网上软件行业的招聘需求和薪资水平， 包括公司概况、学历、工作岗位等数据。并利用数学方法，建立了多元逻辑回归模型，探究了公司规模、学历、工作岗位等因素与薪酬的依赖关系。

2 数据处理与分析

文章的数据来自于国内某招聘网站的的招聘信息，共计 46154 条有关于计算机行业的招聘信息，每一条信息包括：工作岗位、工作地点、公司名称、公司规模、公司类型、学历要求、薪资、福利待遇字段。因为数据很多都是文本结构，所以对数据进行定量化分析，对公司规模、公司类型、学历要求、薪资福利进行了转换，并对薪资进行了平均值的计算。

对公司类型进行了划分，将其分为了'外资','民营','国企','合资','上市','事业'以及未知七类分别标记为 1-7；将公司规模划分为'少于 50 人','50-150 人','150-500 人','500-1000 人','1000-5000 人','5000-10000 人','10000 人以上'7 种分别标记为 1-7 号；将学历要求划分为"中技","中专","高中","大专","本科","硕士","博士",'未知'8 种编号为 1-8；关于工作福利，通过将获取的文本进行分类，以福利待遇的多少来划分，标记为福利待遇的个数。

通过对数据的量化处理，以及去除不全信息等最后保留了 45503 条数据提供给模型训练预测。利用 pandas 工作可以所有数据的详细信息，如表 1 所示。

表一 数据描述

	Company_type	Company_size	Graduation	Jobwelf
Count	45503	45503	45503	45503
Mean	2.38	2.27	3.59	5.31
Std	1.31	2.21	1.19	3.02
Min	1	0	0	1
Max	7	7	7	13

Pearson 相关系数（Pearson Correlation Coefficient）是用来衡量两个数据集合是否在一条线上面，它用来衡量定距变量间的线性关系。文章利用量化的数据求解 Pearson 相关系数，初步判断得到的数据是否对最后的预测有利，初步的结果如表二。

通信作者：高云龙（yunlong@mail.dlut.edu.cn）

表二 各个字段的相关性分析

	company_type	company_size	graduation	jobwelf	salary
company_type	1.00	0.03	0.03	0.02	0.06
company_size	0.03	1.00	0.10	-0.18	-0.01
graduation	0.03	0.10	1.00	-0.03	0.07
jobwelf	0.02	-0.18	-0.03	1.00	0.03
salary	0.06	-0.01	0.07	0.03	1.00

从表二可以看到，公司规模对薪酬的关系性很小，且呈负相关，而且其他因素的相关性也不是特别的高，这有一部分原因是数据在进行量化时处理的不够标准。

3 预测与方法调整

利用 Pytorch 平台搭建的三层网络结构，通过量化的公司类别、学历要求、工作福利作为输入特征，通过三层的神经网络[2]，进行预测工资薪酬，以千为单位，文章进行了 300 次迭代训练，模型的 loss 值趋于稳定，但是预测结果趋于一致在 8k 左右徘徊，效果不是很好，预测结果不够理想的原因可能有数据量化处理比较简单和数据来源比较单一，未进行严格的筛选。在此篇文章暂不列出。

虽然通过全连接的神经网络为达到比较好的预测效果，因为有很大一部分原因是数据处理的过程没能做到标准化，以及数据来源不一定全部都是有效的，需要做进一步筛选。为了进一步继续实验得到更多有用信息，利用提取到的现有数据，进行接下来的实验，稍微改变一下模型分类预测策略，将薪酬划分为 10000 以下标记为 1，10000-20000 标记为 2，20000-50000 标记为 3，50000 以上标记为 4 进行分类预测。

在 1980 年 Ohlson 第一个将逻辑回归方法引入财务危机预警领域，他选择了 1970～1976 年间破产的 105 家公司和 2058 家非破产公司组成的配对样本，分析了样本公司在破产概率区间上的分布以及两类错误和分割点之间的关系，发现公司规模、资本结构、业绩和当前的融资能力进行财务危机的预测准确率达到 96.12%。逻辑回归分析方法使财务预警得到了重大改进，克服了传统判别分析中的许多问题，包括变量属于正态分布的假设以及破产和非破产企业具有同一协方差矩阵的假设。文章想利用逻辑回归的方法，根据公司规模、公司类型、学历要求和福利待遇字段进行逻辑回归训练，预测对应薪酬的等级，如前文所划分的 4 类。通过 sklearn 准确率计算得到准确率的值 0.78，预测分析图（图 1）

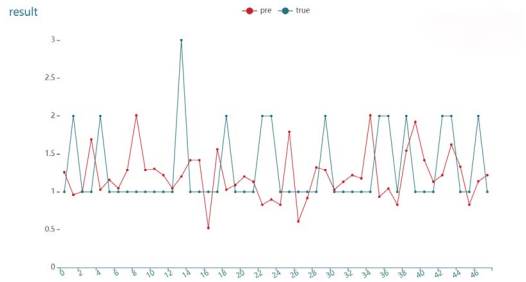

图 1 预测结果分析图

4 结论

本文通过搭建爬虫获取某招聘网站 4 万多条招聘信息，经过一系列的数据清洗处理后，通过数据分析及挖掘得到如下一些结论：（1）公司规模对薪酬的影响不大；（2）通过多元逻辑回归可以基本预测薪酬的等级；（3）IT 行业薪酬普遍过高，对于更好预测薪酬需要引入更多的字段信息。

5 总结与展望

文章利用公司规模、公司类型、学历要求和福利待遇字段信息想预测薪酬，但是因为数据的单一来源，字段的过少，数据处理过程简单等原因没有达到比较理想的结果，后续的研究将引入更多的字段包括，城市，岗位类型，工作经验等更多信息，利用深度学习网络重新搭建模型，实现薪酬的更好预测。

参考文献

[1]. 郭丽清，蓝康伟，朱思霖，李泓锴，许颖. 基于大数据的互联网行业人才薪资影响因素分析. 《计算机时代》2020 年第 2 期 9-12, 17 共 5 页.

[2]. 梅端 周会会. 基于神经网络的城镇在岗职工平均工资预测. 《河北工程技术高等专科学校学报》. 2016 年第 1 期 36-38, 共 3 页

[3]. 加春燕 ， 尚意展 ， 钟宇辉 ， 冯文昊 ， 罗丹. 多元线性回归模型在软件行业薪资影响因素分析中的应用. 《北京工业职业技术学院学报》- 2019 年 3 期