

基于机器学习的推荐与评价方法

黄星寿¹, 刘 迪^{1,2}

(1.河池学院数学与统计学院, 广西 宜州 546300;

2.武汉大学计算机学院, 湖北 武汉 430072)

摘 要: IT类专业学生由于其专业特点, 企业实习环节往往贯穿整个培养过程, 实习环节效果的好坏直接影响到学生的能力培养与就业质量。如何将实习单位的资源配置、业务特点及学生专长与兴趣等因素进行有机整合, 是改善和提高实习效果的有效途径。本文基于机器学习的方法, 对IT专业学生实习单位推荐与评价开展了研究工作, 以某高校计算机专业历年的实习、评价和就业等相关数据为学习样本, 自动学习和生成推荐模型与评价体系。实际应用效果表明: 该系统能为实习组织工作提供更加客观的决策支持信息, 有效提高学生的实习与就业质量。

关键词: 推荐系统; 机器学习; 评价系统

中图分类号: TP181 **文献标识码:** A

A Machine Learning Based Recommendation and Evaluation Method

HUANG Xingshou¹, LIU Di^{1,2}

(1.College of Mathematics and Statistics, Hechi University, Yizhou 546300, China;

2.Computer School of Wuhan University, Wuhan 430072, China)

Abstract: The effect of internship experience is a key factor of ability training and employment for college students, especially for IT majors. It is an effective way to improve the practical effect of business internship by integrating the resources and business characteristics of the firms with the expertise and interests of the students. This paper proposes an internship recommendation and evaluation system for IT major students based on machine learning methods by using college's internship and employment data as sample data to generate the recommendation model and the evaluation system. The findings show that the system can provide more objective decision-support information for the organization of internship, and improve students' internship and employment.

Keywords: recommended system; machine learning; evaluation system

1 引言(Introduction)

国家信息化建设的迫切需求催生了大量IT企业的涌现, 同时也提出了持续性的IT专业人才需求。顺应这种趋势, 几乎所有的高校都开设了IT类专业。在广招生源的同时, 各培养单位也都意识到企业实习环节在IT类专业学生培养过程中的重要性, 也开展了大量有针对性的研究工作。

文献[1]对工科类大学生的成长方式进行了探索与总结, 通过雷达图的形式给出了“实习、实训”和“科技创新训练”环节对学生就业能力培养的影响, 详细的论述了以就业为导向的工科类大学生培养应该尤其注重加强企业实习环节的管理, 充分利用好这一宝贵的社会资源, 以培养出能满足社会需求的专业人才; 文献[2]基于“5S管理理论”分析和论述了加强高校学生实习、实训环节管理工作的必要性, 并给出了相关的实施流程及可行性分析报告; 文献[3]以促进学生

就业为出发点, 详尽的介绍了美国高校对于学生实习的有关价值观念的形成过程和与之相对应的政策制定与演进, 实习过程的组织与效果测评等内容。随后与我国的相关制度与组织过程进行了深入对比, 强调了“制度化”作为“基石”的重要性; 文献[4]也从促进学生就业的角度出发, 探讨了加强实习基地建设, 缓解就业压力和提高就业质量的必要性和紧迫性, 并对研究小组所开展的探索工作进行了简洁的总结与反思; 文献[5]分析了在信息化背景下, 教与学的过程在执行环节中存在的一些问题, 着重强调了“过程”的重要性。而实习环节也是整个过程中极为重要的一个环节。

2 机器学习辅助决策(Decision-Making aided by machine learning)

机器学习是指使用计算机通过模拟人类学习和获取信息的准则, 以预测为目标的一系列过程。它包括统计建模、优

化处理、算法设计和统计分析等,涉及数学、统计学和计算机科学等多个学科^[6]。通过几十年的发展,尤其是随着计算机处理和存储能力的日益提高,机器学习目前已经成为一个热点研究领域,并和各行各业紧密结合,成为一门“利器”。决策论作为运筹学的一个重要分支,为决策分析提供了坚实的理论基础^[7]。传统意义上的决策论往往需要预先给定一个评价准则,随后在给定的信息集上通过数量方法来寻找或选取最优决策。当面临的数据维度较高,组成复杂的时候,常规的数量方法难以挖掘出高维数据中所隐含的特征,导致容易偏离最优决策。这种情况下,可以通过利用新的技术手段进行高维数据挖掘来改善效果;也可以转变角色,退化为辅助决策来继续发挥作用。麻省理工学院的资深学者Theja Tulabandhula和Cynthia Rudin在文献[8]中提出了一种绑定机器学习和决策的框架,并在航线规划和交通路径规划(ML&TRP)等实际应用领域开展了验证性研究工作,在一定程度上证明了该方法的理论基础和实际可行性。文献[9]利用贝叶斯网络对不确定性问题的表达与处理能力,设计了一套网络交互教学效果评价系统,能有效改善网络教学效果评价的质量。

本研究小组对我院计算机科学与技术、软件工程、网络工程等三个专业,自2011年以来的实习与就业相关数据进行收集整理,涉及相关学生累计达785人,实习与就业单位达325家,统计的多维度的单位信息、学生信息、实习组织相关信息、评价与反馈信息等记录高达300多万条(维)。依靠人工已经难以充分和有效的挖掘出这些信息内部所包含的有价值信息,以为后续实习工作提供辅助决策。因此,本文利用机器学习的方法对2011—2014年的相关数据进行学习,从高维度数据中学习出各个实习/就业单位和学生的相关特征,给出明确的类别标识,并作为决策信息为2015届的实习与就业组织工作提供参考依据。

3 辅助决策的推荐系统构建(Construction of recommendation system for aided decision-making)

从宏观上来说,将合适的学生派遣至合适的实习单位,能充分发挥主观能动性和资源配置优势,达到最好的实习效果,从而提升学生的整体就业竞争力。具体到每位学生,只能从众多的待选实习单位中选择一家进行派遣,且在实习过程中进行改派的可操作性也不强。因此,如何准确的对实习单位和待派学生进行特征分析与匹配,成为要解决的关键问题之一,也是首要问题。具有自然属性的“实习单位”和“实习生”完全能符合“物以类聚,人以群分”的属性,如果能借助于机器学习算法对“实习单位”和“实习生”进行合理的分析与划分,将能有效的提高实习派遣与管理的效率和改善效果。

本系统的构建目标即为:利用机器学习算法,以历史的实习与就业数据为样本,对相关实习单位按照多维特征值进

行聚类操作,形成 k 个类;在实习派遣阶段,再利用分类算法,将每名同学分到 k 类中的一个。这样就能建立起一名学生到某类实习单位之间的映射关系,辅助决策推荐系统示意图如图1所示。

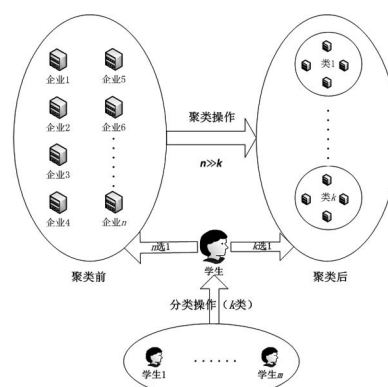


图1 辅助决策推荐系统示意图

Fig.1 Diagram of recommendation system for aided decision-making

在图1中, m 为学生总数, n 为实习单位总数, k 为实习单位总类数,其中, $n \gg k$ 。通过该“聚类—分类”操作,实习派遣操作就转换为从系统为某位同学推荐的一类实习单位中选择一个的问题。该系统要能完成相关功能操作,需要解决如下两项关键问题:

3.1 聚类算法选择

聚类算法目前已经在众多领域得到广泛的应用并取得了良好的效果,尤其是在商业推荐、社团划分等应用领域。在本系统中,聚类算法的目标为:从纷繁芜杂的实习单位相关高维数据中,提取出关键性的特征向量,并以此为依据将所有的实习单位聚为 k 个类。这 k 个类将作为后续分类操作的依据。聚类操作的起源可以追溯至古老的分类学,在计算机的存储和处理能力达到一定水平之前,利用数学工具进行定量的分析存在着困难,人类只能在经验和专业知识的基础上执行带有较强主观色彩的判断。这种方式已经难以适应目前以大数据量为背景的应用场景。与此同时,各种基于机器学习的聚类算法大量涌现并且开始具有实际应用价值。其中,基于划分式的方法发展得到了较多的关注,研究成果也较为丰富。划分法的基本思想为:给定一个包含 N 个元素的数据集,通过分裂的方法将其构造为 k 个分组。为了保障算法快速收敛,一般会要求同一分组中的元素之间的距离满足一个给定的阈值(距离小则认为相似度高)。

(1) K-MEANS算法

K-MEANS算法是一种较为经典的聚类算法,其基本思想为:根据总类别数量 k ,在样本中随机找出 k 个点来作为原始的类中心点,然后计算余下的点与选定的 k 个点的距离,按照距离将其归入某类,完成操作后再重新计算 k 类中所有距离的平均值并将其作为新的中心点,不断的迭代,直到测度函数收敛(中心点不再发生明显变化)。通过该算法,可以将相识

度高的点聚为一类，同时将不同类之间尽量分开。但是，利用K-MEANS算法所划分的类别之间的差异度往往不够大，且对样本数据要求较高，在实际的应用过程中需要采取各种改进措施。

(2)K-MEDOIDS算法

为了克服K-MEANS算法对脏数据敏感的缺点，K-MEDOIDS算法利用一个对象(MEDOIDS)来代替初始中心点，然后进行初始聚类，再找出类中到其他点距离之和最小的点作为新的中心点，再重复该操作直到收敛。Partitioning Around Medoids(PAM)算法是该类算法中具有代表性的一种，但是由于存在较多的循环和迭代运算，算法复杂度较高—— $O(k(n-k)^2)$ 。文献[9]提出了一种简单高效的启发式算法将计算复杂度降低为—— $O(nk)$ ，使得其实用性大幅度提高。

在实际操作过程中，并不会向所有的实习单位都派遣实习生，尤其对于IT企业，单位来源的动态性也较强，即使是同一单位，在不同时期所体现出的对实习生的特征也不尽相同(根据企业实际工程项目)。因此，对实习单位的聚类操作需要动态的进行，以适应实际情况。经过综合对比分析和实验测试，我们选择文献[10]提出的基于K-MEDOIDS的改进算法来作为本系统的聚类算法。

3.2 分类算法选择

分类算法的研究工作也积累了大量的研究成果，可供选择的的面也较为丰富。但本系统所需的分类方法有极强的特性，主要体现为：企业类和实习生不具有同构性，即某类企业的特征与某名学生的特征没有直接的相似性，不能直接按聚合的k类来对学生进行分类。因此，我们需要建立起企业类到实习生之间的关联规则，以实现将某位学生到某类企业直接的关联(推荐依据)。任务转换为“关联规则挖掘问题”，该问题是数据挖掘中的一个重要领域，而基于关联规则的分类方法通常包含两个基本步骤：首先，利用算法从样本数据集中挖掘出所有的满足预先指定支持度和置信度的类关联规则；接下来利用启发式算法从第一步给出的类关联规则中挑选出恰当的规则，用于分类操作。采用在线学习的思想，文献[11]提出了一种如图2所示的模糊关联规则挖掘方法。

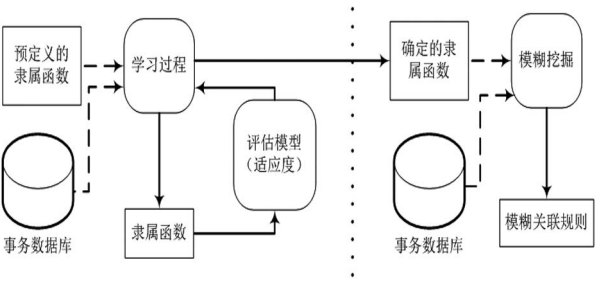


图2 模糊关联规则挖掘方法示意图

Fig.2 Schematic diagram of fuzzy association rules mining

如图2所示的关联规则方法可简述为：首先给出一个预定义的隶属函数，通过学习过程来学习事务数据库中的对象(企业类/实习生数据集)并在线对模型的适应度进行评估，最后给出确定的隶属函数，再利用模糊挖掘方法从数据集中挖掘出模糊关联规则，用于指导分类过程。利用该方法可以有效的解决本研究所涉及的异构对象之间的关联问题，模型评估过程可以作为下一阶段效果评价环节的反馈入口，提高准确率、增强系统的适应性。

4 效果评价模型(Effect evaluation model)

单纯考虑具体的实习成绩评定等细节性环节，往往会导致效果评价环节流于形式，失去评价的作用。本系统将综合考虑实习生反馈、企业反馈、就业情况、教师评价等多方面的因素，力争客观评价，且对实习效果的评价将作为影响因子(λ)反馈至系统的分类模型环节，用于评估、调整隶属模型，更好的支撑模糊关联规则的挖掘过程，从而改进派遣环节的分类效果，提高派遣环节的针对性，最终促进和提高实习效果，实现实习生、学校和实习单位三方的和谐发展。 λ 的构成如表1所示。

表1 λ 构成表

Tab.1 λ composition table

评价主体	评价内容	权值	评价方式	反馈形式
实习生	与预期符合程度	0.2	打分	实习报告
实习单位	是否愿意继续接纳同类实习生	0.2	选项、评语	实习报告
责任教师	实习成绩考核	0.1	打分、写评语	成绩、评语、说明
领导小组	就业情况	0.5	打分	就业协议书、就业合同、走访调查

实习生可以在实习报告中，对实习派遣与预期情况的匹配程度进行打分评价，占0.2的权值；实习单位可以通过实习生的实习报告向学校反馈其是否愿意继续接纳同类实习生，以及给出相关评语等，占0.2权值；责任教师可以结合日常考核来对学生实习期间的表现进行打分和评定，如有必要，也可以给出相关说明，该环节类似于传统的实习成绩评定，占0.1的权值；领导小组负责后续就业相关情况的跟进调查，主要依据是就业协议书、就业合同和走访调查的结果等，占0.5的权值。因为就业情况能比较客观的反映实习派遣的效果，例如，如果某位实习生从系统推荐的一类实习单位中选择一个并最终在该实习单位就业，认为该派遣为一项正确的派遣，故设定较高的权值，有利于隶属函数的优化和关联规则的挖掘。评价模型输出的 λ 值为归一化的标准值，可以直接应用于反馈环节。

5 应用效果与分析(Application effect and analysis)

本小组将系统应用在我院2015届161名毕业生的实习派遣

和效果评价环节,并对效果进行了分析。总的实习单位数量为78个(基本都为网络、软件、培训等IT类企业),实习单位聚类情况如图3所示。

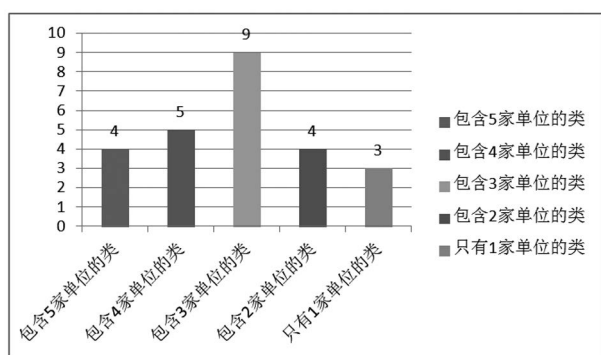


图3 实习单位聚类情况

Fig.3 Clustering of internships

分析图3的数据,我们可以发现,系统将这78家单位聚为25个类,且绝大多数的类包含的单位数量都在2至5家,仅有三个单选类,我们查看原始数据发现其分别为医疗、交通和政府机构等与IT企业相关性不强的单位,因而很难找到共性特征。这样的聚类效果为后续模糊关联和推荐打下了较好的基础。

在所有的161名毕业生中,有五人与非IT类单位有明确的就业意向,其实习派遣直接指定。实际参与推荐的实习生总共为156名,所有的同学都很快的从推荐类中选取了自己认为合适的单位并顺利完成实习过程。目前共确定有效就业人数为155人,占总毕业人数的96.2%,统计数据具有说服力。在应用本系统之前,实习派遣很难有针对性的开展,学生最后的就业也基本与实习单位没有关联。应用本系统后的2015届毕业生中,有87人在实习单位就业,有10人表示期望与实际情况不符合,有16家实习单位表示实习生能力有待提高,将加权统计得到的 λ 值反馈至模糊关联学习模块,系统为8名学生给出了与之前不一样的推荐类。这表明系统在具备稳定性的同时,也能根据实际情况进行自适应调节。

6 结论(Conclusion)

大数据这一名词不仅频繁出现在学术界、政府报告和各类媒体中,它实际上已经渗透到人们日常生活的各个层面。高效且成熟的机器学习算法恰好为我们提供了分析和挖掘大数据背后隐含规律的工具。本文利用机器学习算法,通过对我们近年来积累的大量实习单位、实习生和就业信息数据进行挖掘,学习并构建出了具有实用性的IT专业学生实习单位推荐与评价系统。在通过聚类算法实现对实习单位准确聚类的基础上,利用数据挖掘方法得出实习生与实习单位直接的模糊关联规则,将其作为指导实习派遣的决策依据,实现高效且有针对性的派遣决策。为适应实习单位和实习生的动态

变化特点,系统给出综合模型来对实习效果进行评价,并以此作为反馈因子来指导前述关联规则的挖掘,使系统具备自适应特性。实际应用效果也进一步验证了系统的合理性和可行性。

本系统的推荐部分已经具备了可直接使用的原型系统,但是在评价部分还存在进一步完善之处。比如:反馈环节的原始信息目前基于纸质材料,部分评价主体基于客观因素可能会做出不太客观的评价,后续考虑修应用远程在线式评价系统,实现评价主体的“盲评”,增强客观性。此外,评价权重值分配是否存在完善之处,还有待进一步验证。

参考文献(References)

- [1] 于欣欣,李兆博.工科类大学生成长路径的研究与探索[J].现代教育管理,2015(1):124-128.
- [2] 何瑜.高校学生实训实习管理模式创新研究——基于5S管理理论[J].内蒙古师范大学学报(教育科学版),2015,05:76-77.
- [3] 朱红,凯伦·阿诺德,陈永利.制度的基石、保障与功能——中美大学生实习比较及对就业的启示[J].北京大学教育评论,2012,01:107-123;190.
- [4] 詹一虹,侯顺.加强实习基地建设拓宽高校毕业生就业渠道[J].教育研究,2006,09:90-92.
- [5] 蔡曼君.信息技术环境下“学”与“教”分离现象透视和成因分析[J].电化教育研究,2013,02:93-99.
- [6] Christopher Bishop. Pattern recognition and machine learning [M]. springer, 2007: 138-216.
- [7] Perry J. Williams, Mevin B. Hooten. Combining statistical inference and decisions in ecology[J]. Ecological Applications, 2016, 26(6): 1930-1942.
- [8] Tulabandhula T, Rudin C. On combining machine learning with decision making[J]. Machine Learning, 2014, 97(1-2): 33-64.
- [9] 张晓勇,彭军,文孟飞.基于贝叶斯网络的网络交互教学成效评价系统[J].现代远程教育研究,2012,04:85-90.
- [10] Park H S, Jun C H. A simple and fast algorithm for K-medoids clustering[J]. Expert Systems with Applications, 2009, 36(2): 3336-3341.
- [11] Alcalá-Fdez J, Alcalá R, Gacto M J, et al. Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms[J]. Fuzzy Sets and Systems, 2009, 160(7): 905-921.

作者简介:

黄星寿(1963-),男,本科,副教授.研究领域:统计学,微分方程.

刘迪(1980-),男,博士,副教授.研究领域:机器学习.