

基于 BP 神经网络的就业招聘企业客户分类问题研究

乔 非 葛彦昊

(同济大学电子与信息工程学院 CIMS 中心 上海 200092)

摘 要 在高校就业信息化建设中,对进入高校招聘毕业生的企业客户进行等级分类和预测能够有效帮助高校管理者评估与该企业的合作效用,推动大学生就业信息化服务向更具针对性的个性化推荐方向发展。目前该问题的解决方案大多基于从业人员的主观经验判断,缺乏完善的定量分析模型。抽取某高校教育管理信息系统中与进校招聘企业客户相关的数据样本,并借助 BP 神经网络模型搭建有效的数据分析模型,使用训练样本数据训练模型从而确定分析模型中各节点参数,将训练后的分析模型用于预测测试样本集得到最终的性能指标,最后将该模型分类性能与当前同类问题的其他解决方案进行比较。对比结果显示,基于 BP 神经网络模型分类方式在预测准确度和精度方面明显优于其他现有模型。该研究成果能够在信息化平台中为该问题提供高效的解决方案,帮助高校就业工作管理者及服务人员依靠该模型对进校招聘企业做出快速精准的客户等级预测,为高校就业服务工作决策提供支持。

关键词 BP 神经网络,数据挖掘,分类问题,就业信息服务

中图法分类号 TP301.6 **文献标识码** A

Customer Classification Model of Employers by Using BP Neural Networks

QIAO Fei GE Yan-hao

(CIMS Center, School of Electronic and Information Engineering, Tongji University, Shanghai 200092, China)

Abstract Customer classification of employers brings great benefits for correctly evaluating the category of each employers, which helps decision-makers to appraise the corporation efficiency when they corporate with each customer. In order to give quantitative model to help solving this problem, we extracted raw data series from the educational management information system and built a classification calculating model based on BP neural network. Then we managed to get each value of the parameters in the model by training the historical datasets and compared the result of prediction with other methods which are widely used nowadays to solve this problem. Finally we made a conclusion that customer classification of employers for career guiding service by using BP neural network model performs better than other existing solutions and gives more efficient supports to management layer of universities and government by making comparably precise predictions.

Keywords BP neural network, Data mining, Data classification problem, Career guiding information service

1 引言

近年来,全国高校毕业生人数不断激增,2013 年全国毕业生人数达 699 万人,被社会媒体称为史上“最难就业季”;然而随后 2014 年全国毕业生总数在此基础上再创新高,达到空前的 729 万人。面对如此严峻的就业压力,政府及各高校就业服务部门迫切希望为毕业生提供更好的就业服务工作。若要进一步提升就业服务水平,就要为毕业生提供更加细粒度、有针对性的服务工作,随着校园信息化建设的不断深入,各高校的信息部门逐渐将前沿的信息技术及数据分析技术应用于校园信息化领域,以此更好地服务在校学生,其中就业信息化建设就是校园信息化建设的重要一环。以此为基础,对于校方就业工作管理者而言,当校方举办各类招聘活动或某一招聘企业希望与校方合作招聘时,精准地定位某一进校招聘企

业的客户等级能够有效帮助其评估判断与该企业的合作效用,一方面对于重点企业客户维系关系提供更多支持,另一方面对于以往合作较少的优质新客户加强合作交流,从而进一步提高服务水平,有效支持决策,进而提升在校生就业质量。此外,对于在校大学生而言,了解该企业对于学校的招聘客户等级,能够帮助他们了解该企业对于本校毕业生的重视程度,辅助评估个人进入企业后未来发展的成长空间,帮助毕业生选择最为合适的招聘企业并与其签约。

针对以上的业务需求,目前业界大多采用基于线性决策模型或专家经验的静态方法进行企业客户分类,其判别准确度普遍偏低,缺少相对完善的、高效的定量决策模型支持。这种传统的客户分类方式不仅高度依赖于资深的业务经验,对于缺少经验的从业人员,这种判断方式几乎不可行;而且对于一些新兴的潜在企业客户,如果没有精准的数据模型,即使是

本文受同济大学校园信息化项目资助。

乔 非(1969—),女,博士,研究员,主要研究方向为系统工程优化算法、控制工程以及企业信息化等,E-mail: fqiao@tongji.edu.cn;葛彦昊(1990—),男,硕士,主要研究方向为系统工程应用、数据挖掘和机器学习等。

资深从业人员也容易错误地评估其客户类型。因而需要一种基于定量的数据分析模型对所有进校招聘企业客户进行分类判断,从而解决以上问题。

另一方面,随着误差反向传播学习算法在 20 世纪 80 年代中期由 David Rumelhart、Geoffrey Hinton 等人再次独立提出(1970 年 Paul Werbos 曾首先提出该算法,但当时学术界并无人知晓),人工神经网络(Artificial Neural Network,ANN)这一经典的机器学习模型再次迎来了蓬勃发展时期,目前我们可以将该模型广泛应用于各类数据分类、聚类以及回归的数据挖掘应用领域。其中 BP 网络(Back-Propagation Network,BP)因具有高度的非线性映射能力、结构简单、性能良好,成为 ANN 最为常用的神经网络模型之一。若能够把 BP 神经网络模型引入到就业信息化服务领域,将会很大程度上解决该领域的数据挖掘问题,拥有多种应用场合,例如寻找招聘企业与应聘者的匹配关系、预测毕业生的求职行为以及向招聘与应聘双方推送个性化的信息服务等等。

为达到进一步提升高校就业服务水平的目的,本文将着重讨论对进校招聘企业客户等级的分类问题。为有效解决该问题,首先抽取来自校内教育管理信息平台的招聘企业信息以及其历史招聘行为数据,基于 BP 神经网络的数据分类模型并依据具体业务知识搭建有效的数据分析模型;其次将经过预处理的样本数据作为训练样本,利用提出的模型对样本数据集进行训练和迭代,确定模型中各节点参数值;最后将基于这类模型的分类预测结果与当前同类问题的其他解决方案进行横向对比,以论证本文所提出的模型在企业客户分类预测问题中的分类结果比其他方案更优。

2 基于 BP 神经网络的数据模型搭建

2.1 数据模型及预处理

在各高校的就业工作部门中,工作人员每年通常会将各

类进校招聘企业的相关信息录入到全校的教育管理信息系统中,其中包括该企业的招聘、在校生命明细条目以及各类招聘活动的举办情况等。本课题研究中抽取了总计 1019 家企业用户,并依据专家法将其分为 5 类,如表 1 所列。

表 1 用户等级分类描述

排序	类别	描述
1	新客户	仅保持联系的待拓展客户
2	一般客户	适合作业的客户群体
3	资深客户	推荐合作的客户群体
4	重要客户	长期合作的重要企业客户群体
5	重点客户	急需维系关系并加强合作的重要企业客户群体

由于在本问题研究领域,尚未有学者建立科学可靠的定量模型供分析和预测,因此需要首先对提取的数据进行相关度分析,确定其中相关度最低的数据维数作为输入序列。结合相关业务知识,选取了表 2 所列的七维数据作为有可能选取的输入数据项。

表 2 输入数据属性及描述

属性	描述
公司规模	公司规模大小(由公司在职工工数决定)
是否世界 500 强	该企业是否为世界 500 强
单位性质	招聘单位的注册类型
单位地区	招聘单位的注册地址所在地区
招聘活动举办种类	近 3 年中该企业举办过的招聘活动种类数
招聘毕业生总数	近 3 年中该企业共招聘毕业生总人数
进校招聘年数	近 3 年中该企业进校招聘年数统计

在完成以上数据的抽取工作后,将以上七维数据与最终客户分类结果联接在一张表中。在该模型中,分类结果数据为专家评定值。我们假设以上七维数据与其最终分类结果存在非线性映射关系,表 3 展示了部分待分析数据的输入输出样本,在联接过程中以相同的企业 ID 进行关联操作(为便于统计,将实际企业 ID 进行了代码转换)。

表 3 用于训练的待分析数据样本

企业 ID	公司规模	单位性质	单位地区	是否世界 500 强	招聘活动举办种类	招聘毕业生总数	进校招聘年数	分类结果
1	中	三资企业	上海市	否	0	1	1	1
2	大	三资企业	上海市	否	0	9	2	4
3	中	教育单位	河南省	否	0	10	3	3
4	大	其他企业	上海市	是	0	89	3	5
5	小	国有企业	北京市	否	0	1	1	1
...	...	...	...	...	...	...	...	...

在对样本进行训练之前,如果样本数据不够完整将会导致训练结果偏差甚至错误,因此需要对原始的训练数据序列进行预处理工作。对数据的预处理工作主要包括补填缺失数据、数据清洗以及数据的归一化处理等。其中补填缺失数据操作的原则是向下填充,即当前空缺值的补填值取上一条记录中该字段的取值,由于原始数据集排列顺序随机,因此该补填方式不会影响整体数据集在模型训练和预测过程中的有效性和随机性。

在实际的数据抽取过程中,最先获得的数据维数众多,需要清除其中的冗余数据。在数据清洗过程中首先进行数据相关性分析,保留相关系数较小的数据序列作为样本。采用如下公式计算数据项之间的相关系数:

$$Correl(X,Y)=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2(y-\bar{y})^2}}$$
 (1)

其中, $\bar{x}$ 与 $\bar{y}$ 分别为 x 与 y 的平均值。

各项数据列相关系数的计算结果如表 4 所列。

表 4 输入数据序列的相关性分析

训练数据相关度	单位标签	公司资本	招聘活动类型统计	招聘总人数	参加招聘年数	单位性质	单位地区
单位标签	\	0.14	0.12	0.02	0.07	0.10	-0.03
公司资本	0.14	\	0.07	0.12	0.14	-0.07	-0.10
招聘活动类型统计	0.12	0.07	\	-0.06	-0.29	-0.06	-0.01
招聘总人数	0.02	0.12	-0.06	\	0.36	-0.08	0.04
参加招聘年数	0.07	0.14	-0.29	0.36	\	-0.09	-0.02
单位性质	0.10	-0.07	-0.06	-0.08	-0.09	\	0.11
单位地区	-0.03	-0.10	-0.01	0.04	-0.02	0.11	\

由以上结果比较可以得出,“招聘总人数”与“参加招聘年数”指标的数据相关系数最大(相关系数为 0.36),因此决定在训练样本前舍去其中一维数据,又由于“参加招聘年数”这一指标与其他各指标的平均相关系数大于另一候选指标,因此从训练样本数据中舍去“参加招聘年数”这一指标及其相关数据,保留剩余的六维输入数据。

最后,需要对保留的数据集进行归一化处理。对于数据的归一化处理,一般情况下采用以下的公式进行:

$$y = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

其中, x, y 分别对应为归一化处理之后的输入与输出值, $x_{\max}, x_{\min}$ 为样本数据的最大、最小值,这里截取其中一维数据作为示例,如表 5 所列。

表 5 原始训练样本数据的归一化处理

公司规模	代码	归一化数值
大	3	1
中	2	0.5
小	1	0

2.2 搭建分类模型

当拥有了完整的数据模型之后,则需要设计合理的分类训练模型。本研究中决定借助 BP 神经网络模型来搭建训练分类模型,该模型是一种多层前馈神经网络,包括输入层、输出层和一个隐含层,各神经元与下一层所有的神经元连接,同层神经元之间无连接。使用 BP 神经网络进行分类预测的基本原理是采用梯度下降法调整各神经元节点函数的权值和阈值,使得网络的实际输出值和期望输出值的均方误差值最小,以此达到分类预测的目的。

之所以借助该模型来解决问题,是因为人工神经网络作为众所周知的生物启发式算法模型,能够有效建模各类非线性问题,而其中 BP 神经网络模型对于建模分类预测问题具有优越的表现性能,且它在解决非线性回归问题时拥有很强的鲁棒性。

图 1 展示了所设计的基于 BP 神经网络分析模型的大致模型结构。

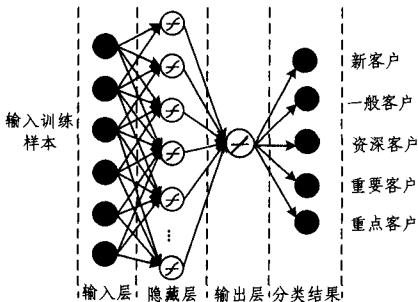


图 1 基于 BP 神经网络的客户分类模型

由于通常情况下,包含输入层、隐藏层和输出层的三层 BP 神经网络模型结构能够被用于回归任意系统,因此采用一个具有该层次结构的 BP 神经网络模型来解决本文问题。其中输入层包括了六维数据样本,分别对应以上提到的六维输入数据序列,以 X_i 来命名输入向量,则:

$$X_i = \{x_1, x_2, x_3, x_4, x_5, x_6\} \tag{3}$$

其中, x_1 至 x_6 代表了上一节中提到的六维输入数据属性。

在隐藏层的设计过程中,选取了 13 个隐藏的非线性神经元进行布局,其中每一个神经元的非线性函数选择为 Sigmoid

函数:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

一般情况下,隐藏层神经元个数的选择原则依据具体问题的不同可以相当复杂,这里通过以下的公式来确定隐藏层神经元个数:

$$H = \sqrt{n_i + m_o} + l \tag{5}$$

其中, H 为隐藏层神经元个数, n_i 为输入节点数, m_o 为输出节点数, l 为 1 到 10 之间的常量。通过以上公式求出的神经元个数 H 即可作为训练时的合理参数。最终为达到最优的性能以及时间的高效性,在进行多次尝试后决定选择 $H = 13$ 。

对于输出层,由于计划分类的数据只有一维,因此输出层包含一个线性神经元。

3 模型性能测试和结果分析

3.1 实验结果记录

当通过以上分析过程确定数据模型及训练模型之后,需要对训练过程中的其他具体参数进行配置,其中的学习率与动量因子参数值为经验值。

表 6 训练过程中参数的选取

设定参数	取值
结果显示步数	20
学习率	0.5
动量因子	0.95
最大训练步数	4000
训练目标	0.005

在训练过程中,首先采用最为常用的最大梯度下降法(简称 LM 训练方法)进行训练。将抽取到的 1019 组数据分为两组,700 条数据用于训练模型,其余 319 条用于预测验证训练后的模型性能。考虑到人工神经网络模型在训练过程中的随机特性,共进行了 10 次训练,记录每一次的训练结果并计算平均值,最终的训练结果如表 7 所列。

表 7 训练结果统计(采用 LM 训练法)

平均训练准确率	平均预测准确率	最大预测误差
608.4/700	237.6/319	2.60

3.2 提高分类模型的泛化性

在解决实际的分类问题过程中,常会因为过度追求训练模型的精度指标而出现“过度训练”的问题,从而导致训练后的模型在处理新的分类预测任务时准确度不增反降,即由于“过度训练”造成系统泛化性降低。为提高分类模型的泛化性,尝试采用贝叶斯正则化方法(简称 BR 训练方法)改进现有 BP 神经网络模型的泛化性能,即引入缓冲系数 γ 使得修正后的网络均方误差变为:

$$MSE_{reg} = \gamma \cdot MSE + (1 - \gamma)MSW \tag{6}$$

其中, γ 为引入的缓冲系数,一般取值范围为 0 到 1,此处取经验值 0.5; MSW 表示各神经元权值的平方和平均值:

$$MSW = \frac{1}{n} \cdot \sum_{j=1}^n W_j^2 \tag{7}$$

由于引入了缓冲系数 γ ,因此系统的 MSE 指标相比最大梯度下降法略有下降,训练速度也相对 LM 训练方法更慢,但是以此换来的是后续预测数据准确性的提升。实验结果如表 8 所列。

表 8 训练结果统计(采用 BR 训练方法)

平均训练准确率	平均预测准确率	最大预测误差
590/700	243.6/319	2.52

对比以上两种训练方法,可以看出它们在解决本问题过程中都拥有各自的表现特点,其中 LM 训练方法拥有较快的训练收敛速率,而 BR 训练方法拥有更优的分类预测性能,其具体对比情况如表 9 所列。

表 9 LM 训练方法与 BR 训练方法的性能比较

训练方法	LM 训练方法	BR 训练方法
训练收敛速率	较快	较慢
分类预测性能	优于 BR 训练方法	略低于 LM 训练方法
陷入局部最优的概率	相对 BR 训练方法更易陷入局部最优	相比 LM 训练方法可有效避免陷入局部最优
泛化性能表现	基本满足需求	优于 LM 训练方法

3.3 与其他分类模型的对比分析

在对已提出模型进行性能测试之后,本小节中将讨论相较于其他分类方法,本文提出的基于 BP 神经网络的分类模型性能是否更优。为回答这一问题,同时实现了对于处理同一数据样本的其他分类算法模拟,其中将目前业内普遍采用的基于业务经验的专家决策方法抽象为一种线性决策模型,同时引入主流的分类算法即朴素贝叶斯分类器作为比较。采用以上各类算法训练和预测的结果统计如表 10 所列。

表 10 BP 神经网络模型与其他分类算法的性能对比

分类方法	预测集正确比例	最大预测误差比例
BP 神经网络(LM 训练)	237.6/319	2.60/4.00
BP 神经网络(BR 训练)	243.6/319	2.52/4.00
线性决策分析模型	166/319	2.00/4.00
朴素贝叶斯分类器	210/319	3.00/4.00

由以上的横向对比结果可以看出,尽管线性决策分析模型在最大预测误差比例(2.00/4.00)指标上相比本文所提出的 BP 神经网络(BR 训练)模型(2.52/4.00)表现更优,但其预测准确率明显低于 BP 神经网络(BR 训练)。相比之下,以 BP 神经网络搭建的分类模型结合贝叶斯正则化训练方法在本问题研究中获得了相对最优的预测表现,预测集正确率达到 76.36%,最大预测误差为 63%,相比于目前主流的线性决策方法在预测过程中正确率提升 10.53%,同时最大预测误差降低了 12%。相较于业内现有模型,本文模型的分类性能明显更优。

结束语 在本文研究中,我们成功搭建了一种基于 BP 神经网络模型的数据分析模型,旨在依据教育管理信息系统中企业的自身信息以及其相关历史招聘数据对招聘企业客户等级进行分类。为达到这一目标,首先抽取相关的源头数据并对其进行预处理以提升数据质量,然后权衡模型复杂度与表现时效性,选择了相对有效的模型结构并设定合理的训练目标参数;在完成训练任务之后,使用训练后的系统对一批未参加训练的输入样本进行预测,记录并计算分类器的正确率和最大误差比例。最后,将设计完成的分类器与现有的其他分类算法性能进行对比,通过对比分析看出基于 BP 神经网络模型的分类器在解决本课题中的分类问题时拥有最优的性能表现。

由于先前的研究中还未有其他学者针对该类型数据搭建有效的数据模型,本课题中的分析模型的有效性需要更多业务实践来印证。与此同时,在搭建 BP 神经网络模型结构时为追求最优的训练表现,采用统计相同的迭代步数下最终的误差函数指标情况,并未考虑相同的处理时间下系统实际的

性能表现,尤其是当后续客户数不断增多时,系统的训练时效性将成为系统整体性能的重要指标参数。此外,如何将 BP 神经网络模型与其他数据挖掘算法相结合亦未在本次研究中探讨,这将作为我们后续研究的主要方向。

参 考 文 献

[1] Baker R S. Educational Data Mining: An Advance for Intelligent Systems in Education[J]. IEEE Intelligent Systems, 2014, 29 (3):1541-1672

[2] Peña-Ayala A. Educational data mining: A survey and a data mining-based analysis of recent works[J]. Expert Systems with Applications, 2014, 41(4):1432-1462

[3] Sen B, Uçar E. Predicting and analyzing secondary education placement-test scores: A data mining approach[J]. Expert Systems with Applications, 2012, 39(10):9468-9476

[4] Borkar S, Rajeswari K. Predicting Students Academic Performance Using Educational Data Mining[J]. International Journal of Computer Science and Mobile Computing, 2013, 2(7):273-279

[5] Qin Yu-Quan, Li Hai-Min. Sales Forecast Based on BP Neural Network[C]//IEEE 3rd International Conference on Communication Software and Networks, 2011:186-189

[6] Wang Rui, Wang Xu. Pulse frequency classification based on BP neural network[J]. Journal of Harbin Engineering University, 2006, 27(suppl.):471-473

[7] Tian Guo-yu, Huang Hai-yang. Identify of hidden layer in Neural network[C]//Information Technology, 2010(10):79-81

[8] Xu Xin, Xu Li-hong. Coal requirement prediction using BP neural network[C]//International Conference on E-Business and E-Government(ICEE). 2010

[9] Hsiao W-H, Chang T-S. Selection criteria of recruitment for information systems employees: Using the analytic hierarchy process (AHP) method[J]. African Journal of Business Management 2011, 5(15):6201-6209

[10] Bakaev M, Avdeenko T. Data Extraction for Decision-Support Systems; Application in Labour Market Monitoring and Analysis[J]. International Journal of e-Education, e-Business, e-Management and e-Learning, 2014, 4(1):23-27

[11] 陈笑怡. 泛在学习中教学质量评价的数据挖掘研究[D]. 上海: 上海交通大学现代远程教育研究中心, 2011

[12] Shang Xiao, Chisholm L A. Classification of Australian Native Forest Species Using Hyperspectral Remote Sensing and Machine-Learning Classification Algorithms[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2013, 7(6):2481-2489

[13] Luan Jing. Data Mining and Its Applications in Higher Education[M]//New Directions for Institutional Research, Spring, 2002

[14] Wanga Feng-hsu, Shaob Hsiu-mei. Effective personalized recommendation based on time-framed navigation clustering and association mining[J]. Expert Systems with Applications, 2004, 27 (3):365-377

[15] Sembiring S, Zarlis M. Prediction of student academic performance by an application of data mining techniques[C]//International Conference on Management and Artificial Intelligence (IPEDR), 2011

[16] 翼正强. 基于 Web 数据分析的就业信息服务平台的设计实现[D]. 山东: 山东大学软件学院, 2013

[17] 吕守涛. 数据挖掘技术在毕业生就业工作中的应用研究[D]. 成都: 电子科技大学, 2007