基于旅游微观数据平台的旅游消费类型预测模型研究

王赛兰[1,2],杨振之[1]

(1. 四川大学 旅游学院,成都 610065;2. 四川大学锦城学院,成都 611731)

摘要:利用旅游微观数据平台中获取的大量旅游消费类数据,建立了一个旅游者消费类型预测模型,在部分缺失旅游消费数据的情况下可以对旅游者的消费类型进行预测和判断。该模型基于监督性学习理论,首先针对已有的完整的消费数据进行学习,使用学习算法不断降低模型的判断误差,直到可以进行比较准确的数据预测;再根据数据缺失情况的不同,采用 BP 神经网络和均值插补的方式进行补足;然后通过 K-means 聚类分析方法,对已经补足的数据进行聚类,从而达到预测判断旅游者消费类型和层次的效果,进而达到在已知部分旅游数据的情况下能对旅游者的消费类型进行预测判断的效果。

关键词:旅游消费;旅游微观数据平台;旅游消费类型预测模型;监督性学习理论

中图分类号:F592.3　**文献标志码**:A　**文章编号**:1000-5315(2017)01-0067-08

消费行为一直是国外消费经济学领域研究的热点问题,学术界广泛认为消费行为随消费者个体特征不同而存在着差异,对不同群体消费行为的类型和差异研究,可以使营销策略的制定和实施更具有针对性。然而,由于微观数据的匮乏,国内关于消费行为研究的成果并不多见,对不同群体消费行为进行研究的成果更为鲜见[1]1。旅游行业中旅游者的个体消费问题研究也存在微观数据缺乏的问题,所以成果较少。本文的前期研究已经建立了一个基于移动互联网以获取旅游微观数据为核心的数字化平台[2]。从旅游微观数据平台上线至 2016 年 10 月 10 日,收集的数据包括游客社会学统计数据、游客时空数据和旅游消费数据。相对于其他两种数据而言,旅游消费数据比较完整系统,实时性高,所以本文选择旅游消费数据作为研究对象。经过对这些数据进行归纳整理,去除一些残缺度高的数据之后,还余下 3 万余组较为完整的数据。旅游消费具有前置性消费特征。考虑到前置消费分析对于游客服务推荐、消费喜好挖掘等具有重要意义,本文着力于建立一个游客消费类型预测模型,在已知旅游者的部分消费,特别是前置消费的情况下,实现其消费水平和消费类型的预测。

在旅游微观数据平台中,旅游消费的原始数据特征设计参考了国家旅游局统计标准,本文系统提取得到的数据可以分为十个维度,分别是长途交通、住宿、餐饮、景区游览、娱乐、购物、市内交通、邮电通

收稿日期:2016-10-12

基金项目:四川省教育厅人文社科研究项目"面向大数据的旅游微观数据信息平台研究"(15SB0323)。

作者简介:王赛兰(1981—),女,湖北武汉人,四川大学旅游学院博士生,四川大学锦城学院副教授,研究方向为智慧旅游、文化遗产与旅游开发;

杨振之(1965—),男,重庆人,博士,四川大学旅游学院教授、博士生导师,主要从事休闲与旅游规划、旅游目的地管理等理论与实践。

讯、旅游天数和旅行人数。

根据已有数据分析,本文通过数据清洗、缺失值插补和聚类方法将旅客聚集到不同的簇中,从而实现游客消费类型预测。本文算法的主要流程为:首先采集游客的前置消费数据,对于存在缺失的数据,根据游客缺失数据量的多少,将采取不同的方法进行处理;然后通过聚类方法形成不同的游客簇;最后实现游客消费层次的预测。

一 旅游消费前期研究成果

从20世纪70年代国外就开始对旅游消费进行了研究,并积累了丰硕的成果,总的来说旅游消费研究中有三个比较重要的模型:需要—动机—行为模型(Gilbert,1991),旅游消费者购买过程模型(Mathieson and Wall,1982),旅游消费者行为的刺激—反应模型(Middleton,1994,2001)。这三个模型研究的角度不同,但都是基于已有的消费数据,而并非预测性研究。

近年来,如何拓展旅游市场成为旅游学者关注的焦点。而要拓展旅游市场,就离不开对旅游消费问题的研究。从本文的研究来看,国内外学者对旅游消费的研究主要集中在消费结构、旅游者消费行为、消费对旅游目的影响等方面。

在消费结构方面,周文丽等从边际消费倾向、消费投向、收入弹性及价格弹性四个方面对我国城乡居民国内旅游消费中的食、住、行、游、购、娱等9类消费的结构进行了实证分析[3]。王媛等以南京市为例,针对具体区域的国内旅游消费结构现状及存在的问题进行定性和定量分析,找出影响国内旅游消费结构的因素,并提出对策[4]。还有针对四川省[5]、河南省[6]或者其他具体区域的类似研究。

消费行为方面的研究主要包括旅游者对旅游决策行为过程、消费行为的完整过程、消费心理和消费满意度的研究。Fang Meng 等从行为学上对游客旅游消费是冲动、计划还是体验进行了研究[7]。有学者对影响旅游消费内、外影响因素进行研究,认为内因包括旅游者自身行为动机、文化背景、旅游偏好等,而外因则包括来自广告的刺激、家庭和朋友圈的影响等等。如在对旅游中的奢侈品消费行为研究中,Park 等针对奢侈品旅游消费从内外因两个方面对其进行了研究,外因有奢侈品的购物场所、奢侈品本身的影响力等等,而内因则是游客本身对待奢侈品不同的态度[8]。还有针对旅游者本身情况的研究,包括旅游者本身经济能力、旅游经验、文化背景、心理特征、年龄性别等等。Myung-Ja Kim 等基于性别差异对韩国网络旅游消费影响问题进行了研究[9]。

在旅游消费与旅游目的地之间的关系研究方面,旅游消费、旅游购物是很多旅游目的地能够成功吸引旅游者的重要因素。Henderson 等研究了旅游购物对新加坡作为旅游目的地的影响,对新加坡作为旅游目的地的成功经验进行了总结,也通过调查分析,对零售业和旅游业中的经验教训进行了反思[10]。还有针对迪拜进行的类似研究[11]。

综上所述,无论是哪一种类型的研究,多是以消费结果作为研究基础,而针对旅游消费个体的相关研究还比较少见。所以,本文计划运用旅游微观数据平台已经获取的大量数据以及机器学习的方法,建立一个对旅游者消费类型预测判断的模型,在部分数据缺失的状况下对旅游者的消费类型进行预测。

二 缺失数据预测

目前针对缺失数据的预测方法主要有均值插补、中位数插补、多值插补以及回归预测法。针对本文提出的问题,我们提出回归预测法和插值填补法相结合的缺失数据填补法。即根据数据缺失的情况,采用不同的数据预测方法,实现缺失数值的插补。

(一)数据预处理

针对数据项:旅行人数、天数、长途交通、住宿、餐饮、景区游览、购物、娱乐、邮电通讯、市内交通费用,可以对后八项数据求取人均每天的消费金额,从而形成本文所需的消费特征。

(二)当缺失数据仅为一项

当原始数据项缺失仅为一项时,本文通过已有的数据项建立回归分析预测模型,即通过游客的大部分消费数据预测缺失待插补的数据,并运用神经网络算法预测缺失数据,从而进行填补。

当原始数据中某项数据成为缺失量时,可以将其作为因变量,其他已知数据作为自变量。我们不妨假设存在这样一个模型:$h_\theta(x_z,x_2,\cdots,x_n)$,能够反映这些随机量之间的关系。于是,可以建立相应的回归模型,如公式2.1所示:

$$y=h_\theta(x_1,x_2,\cdots,x_n)+\varepsilon \qquad (2.1)$$

其中,ε 为误差项,y 表示缺失量的值,$x_1,x_2,$

…，x_n 表示各个非缺失变量，此模型即为多元回归模型。

我们采用神经网络算法，通过训练完整数据集来得到相应的模型 h，在网络权重收敛稳定之后，含有缺失数据的特征向量也就得到确定。BP（back propagation）神经网络是目前应用最广泛的一种神经网络。它是一种基于误差反向传播的多层前馈网络，由信息的正向传播和误差的反向传播两部分组成，是一个包括输入层、隐含层以及输出层的神经网络[12]。

BP 神经网络按监督性学习方式完成训练。当预期输出和实际输出不符时，误差按照梯度下降的方式由输出层经过各个隐含层，最终返回至输入层，以达到逐层修正各个连接权重的目的，因此被称为“误差逆传播算法”[13]28-29。

为了使 BP 神经网络确保样品实际输出与预期输出之间存在的误差能够被控制在较小区间之内，需要对不同层之间的连接权重以及节点阈值做出调整[14]。一般地，将 BP 网络算法的学习过程描述为以下步骤：首先将原始数据进行数据预处理，去除重复、残缺数据，然后通过建立神经网络模型，对完整数据集进行网络训练、仿真，最终通过仿真结果与样本进行对比，从而确定神经网络模型[13]32。

（三）当缺失数据为多项时

当特征数据中存在大量缺失值时，不易对缺失数据进行预测。因此，针对多项数据缺失的情况，可以采用以下方式：将关联性较强的数据进行人工神经网络的预测填补，而对其余关联性较弱的数据可以采用均值插补的方式进行。可以设计如下算法流程：

（1）输入原始数据并进行数据的预处理，去除不一致、重复、含噪声的无效数据；

（2）对数据进行归一化处理，在数据分析之前，通常需要先将数据进行标准化处理，利用标准化后的数据进行数据分析；

（3）网络训练，通过不断调节权值，使网络的输出与预期值相符；

（4）对训练后的网络进行数据仿真；

（5）将仿真结果与样本进行对比，检查数据的拟合度；

（6）根据收敛情况，确定学习类型。

三　旅游消费数据聚类分析

（一）确定聚类特征量及所属类别

原始数据中，长途交通是一个比较重要的特征量；而数据中旅行天数和旅行人数是不同的，无法直接按照数据进行聚类，而且无法得到以“元/公里”为单位的长途消费数据。因此，按照各项消费类目占据总消费的比例处理数据比较合适。所以，首先需要计算得到各项消费类目占据总消费的比例，然后根据消费比例进行聚类分析。

根据消费的重要性，旅游消费分为基本旅游消费和非基本旅游消费两类，其中基本旅游消费中“食、住、行、游、购、娱”尤为重要。在本次数据获取的八个类别数据中，长途交通、住宿、餐饮消费在一次旅游活动中是不可缺少的消费，属于基本消费；而景区游览、娱乐、购物、市内交通、邮电通讯则是弹性的，是选择性比较大的消费类别，属于非基本消费。

基本旅游消费支出可以较明显地反映出旅行者的消费情况水平。通过分析基本旅游消费支出，最终可以得到：在总的旅游消费中，基本旅游消费支出所占比例越高，该旅行者消费水平越低。

由于基本旅游消费支出（Basic）这一特征量是长途交通、住宿、餐饮这三项消费占据总消费的比例，是衡量消费层次的主要的标准，可以直接表征基本消费，所以将其作为一个特征量。第二个特征量为景区游览、娱乐、购物三项消费之和占总消费的比例，用来表征购买力（Purchase）。因为首先这三个类目在旅行过程中比较普遍，选择的自由性程度比较大，所以用来表征购买力是合理的；其次，购买力也是衡量消费层次的另一个重要的特征量。第三个特征量是邮电通讯（Phone），第四个特征量是市内交通（Short_trans）。这两项消费类目比较边缘，将单独列出，由此形成数据分量（表 1 所示）。通过对数据进行归一化处理，可以得到新的特征（表 2 所示）。

表 1.各变量所属类别表

类别	变量
基本旅游消费支出（Basic）	长途交通、住宿、餐饮
购买力（Purchase）	景区游览、娱乐、购物
邮电通讯（Phone）	邮电通讯
市内交通（Short_trans）	市内交通

在 Matlab 中绘制横轴为 Basic、纵轴为 Purchase 的散点图，发现散点分布较为集中，证明这一

组变量的关联性较强,并且可看出基本旅游消费支出越高,购买力越低。而将其他变量两两组合后发现,其余的两两变量的组合都是散乱分布的,并没有什么明显的关系。所以我们可以 Basic 和 Purchase 为标准,划分旅游时消费者的层次,消费者基本旅游消费越低,购买力越高,消费层次越高。而 Phone 和 Short_trans 可以表征消费习惯,而对于消费层次的划定意义不大。

因此,从游客消费数据中得到基本旅游消费支出(Basic)、购买力(Purchase)、邮电通讯(Phone)、市内交通(Short_trans)四个特征量,然后根据特征量再进行聚类分析。表 2 中的数值为每个特征量在总消费数据中占据的比例。

表 2.特征值提取后数据形式

序号	Basic	Purchase	Phone	Short_trans
1	0.5548	0.3713	0.0138	0.0601
2	0.5477	0.3471	0.0145	0.0907
3	0.8992	0.0792	0.0041	0.0175
4	0.8676	0.1117	0.0045	0.0162
5	0.6030	0.3467	0.0000	0.0503
6	0.5748	0.3425	0.0131	0.0696
7	0.7110	0.2642	0.0017	0.0230

(二)聚类过程及结果

聚类分析的目的是将数据划分到不同的簇中。首先,平台能够获得用户记录的游客消费数据,如果用户的消费数据存在缺失项,则根据上节介绍的缺失数据预测方法进行预测,得到预测值后用于填充缺失数据;然后,根据所得到的完整数据与几个聚类中心的距离,确定用户所属的消费层次。下面将使用 k-means 算法对预处理过的数据进行聚类分析,从而得到聚类中心。k-means 算法将 n 个向量 $x_i(i=1,2,\cdots,n)$ 划分成 c 个簇,计算每个簇的聚类中心,确保非相似性指标的价值函数能够控制在最低值。k-means 聚类算法目标函数为:

$$J_m=\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^{m}\parallel x_i-c_j\parallel^2 \qquad (3.1)$$

式 3.1 中,J_m 是目标函数,m 是大于 1 的实数,u_{ij} 是 x_i 属于类别 j 的隶属度(0 或 1),x_i 是第 i 个测量到的数据,c_j 是类 j 的聚类中心,$\parallel * \parallel$ 表示任一测量数据与聚类中心的相似度。

通过下列两式的更新迭代来使上述目标函数达到最小:

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left[\frac{\parallel x_i-c_j\parallel}{\parallel x_i-c_k\parallel}\right]^{\frac{2}{m-1}}} \qquad (3.2)$$

$$c_j=\frac{\sum_{i=1}^{n}u_{ij}^{m}\cdot x_i}{\sum_{i=1}^{n}u_{ij}^{m}} \qquad (3.3)$$

利用 Matlab 中的 k-means 算法的相关函数,得到的聚类中心结果如表 3 所示。

表 3.算法得到的聚类中心

类别	Basic	Purchase	Phone	Short_trans
第一类	0.5415	0.3763	0.0074	0.0749
第二类	0.6003	0.3392	0.0075	0.0529
第三类	0.6571	0.2942	0.0057	0.0430
第四类	0.7229	0.2308	0.0035	0.0429
第五类	0.8745	0.1027	0.0044	0.0184

四　实验计算

(一)缺失值预测模型实验

针对八项预处理后的数据,根据上述神经网络算法,分别对八个数据依次取为缺失项,利用 BP 神经网络算法进行求解。下面以景区游览消费费用为缺失项的例子进行说明。图 1 表示训练参数变化过程也就是计算过程;图 2 表示通过不断训练残差值的变化,可以看出当景区游览消费费用为缺失项时曲线显示出较好的收敛性,证明算法有效;图 3 表示样本点在高维空间基于模型的拟合情况。

同理,当住宿、餐饮、购物、娱乐数据缺失时,通过 BP 神经网络拟合得到的结果,与图 1－3 类似,均具有较好的收敛效果。而当长途交通、邮电通讯、市内交通费用数据缺失时,BP 神经网络训练所得的经验误差无法收敛,说明回归预测处理该问题并不合适,选择用均值差补的方法进行缺失数据填补。

本文通过交叉验证法,先将数据集 D 划分为 10 个大小相等的互斥子集,每个子集都尽可能保持数据分布的一致性,即从 D 中通过分层采样得到。然后,每次用 9 个子集的并集作为训练集,余下的那个子集作为测试集,这样就可以获得 10 组训练/测试集,从而可以完成“k－折交叉验证”。本文定义误差在 10%以内,即可认为缺失数据填补合理有效。表 4 为在测试集中各个数据填补的正确率。

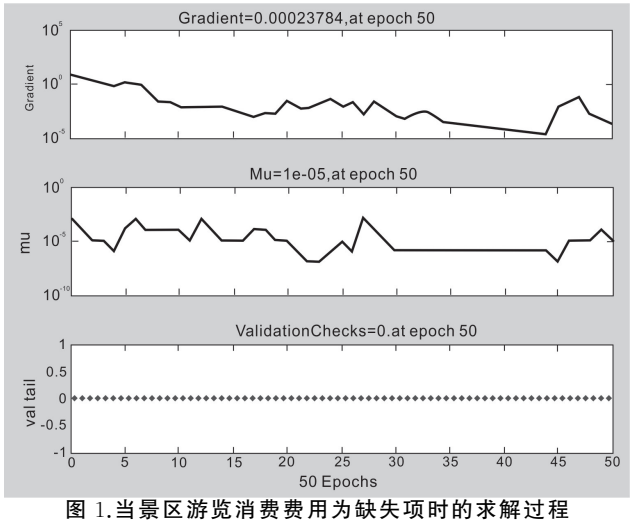

图 1.当景区游览消费费用为缺失项时的求解过程

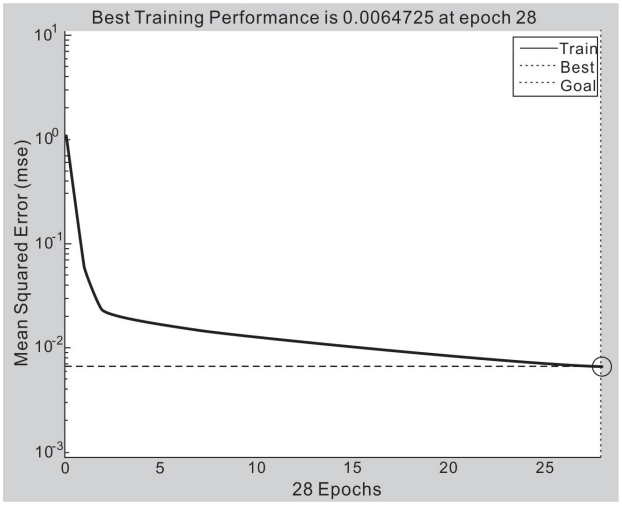

图 2.经验误差变化

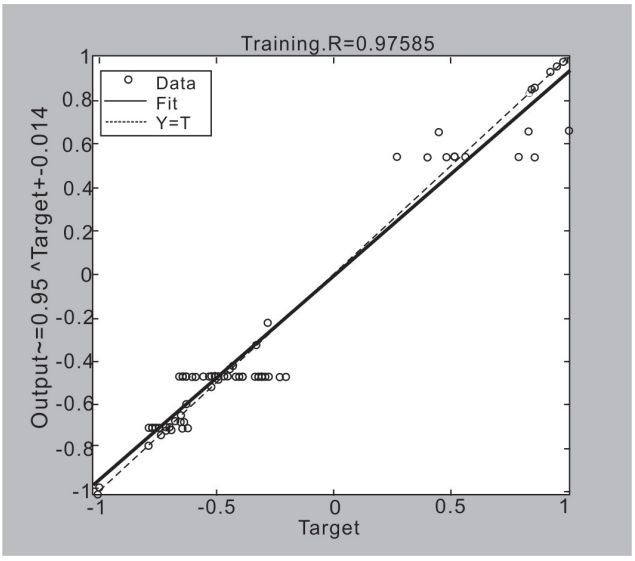

图 3.拟合结果

表 4.测试集中各数据填补的正确率

缺失数据	长途交通	住宿	餐饮	购物	娱乐	市内通讯	景区游览
预测准确率	33.5%	86.8%	90.2%	87.4%	93.4%	54.3%	85.6%

Basic:0.67

Purchase:0.2769

Phone:0.0057

Short_trans:0.0474

（二）聚类分析

为了确定各层次游客的消费水平和特征，我们需要与游客消费水平的平均数据进行比较，以得到较为合理的分析。下面为我们得到的游客消费水平的平均数据。

在用户的旅游消费数据中，平均的基本旅游消费为总消费的 67%，购物消费比例占据 27.69%，邮电通讯占 0.57%，市内交通占据 4.47%。把这个数据作为旅游消费比例的标准，可以衡量得到其他游客的消费层次和消费水平。在对各层次消费结构进行比对中，基本消费是比对的重点，该项占的比例越高，说明该层次的消费能力越低；而购买力项与消费能力呈正向关系，也就是说，购买力占有比例越大游客消费能力越强。根据基本旅游消费和购买力可以直接分出五个消费层次。将表 3 中各层次聚类中心与平均数据进行比对，可以获得各层次的消费特征。

第一类：高消费旅游者

这个消费层次属于整体中消费水平最高的层次，基本旅游消费占到总消费的 54.15%，在基本消费的平均水平中是较低的。购买力这一特征量达到了 37.63%，高于平均水平 9.94%。从整体来说，这一层次的消费者消费能力很强。

第二类：较高消费旅游者

本消费层次属于整体中消费水平较高的，基本旅游消费占到总消费的 60.03%，低于平均水平 6.97%；购买力这一特征量达到了 33.92%，高于平均水平 6.23%，购买能力较强。

第三类：中等消费旅游者

这个层次的消费者消费水平属于中等偏下。基本旅游消费为 65.71%，比平均水平 67%的标准低 1.29%；购买力达到 29.42%，低于平均水平 1.73%。这一层次的消费者市内通讯占到 0.57%，市内交通占据 4.30%，均略低于平均水平。

第四类:偏低消费旅游者

本层次消费者基本旅游消费达到了72.29%,高于平均水平将近5%,购买力低于平均4.61%,消费水平较低。邮电通讯与市内交通也略低于平均水平。整体消费相较于平均水平,属于偏低。

第五类:低消费旅游者

这一消费层次属于消费水平最低的,基本旅游消费为87.45%,高于平均水平20%;购买力低于平均17%。其消费基本针对长途交通、住宿和餐饮。邮电通讯与市内交通略低于平均水平,表明消费水平很低。

(三)实例分析

下面给出两个实例来说明预测游客消费层次的整个过程。例如,用户在行程规划中提前提交了旅行人数、旅行天数、长途交通费和预订住宿费用四项数据,具体如表5所示。

表5.已有的部分数据

类别	实例1	实例2	类别	实例1	实例2
旅行人数/人	8	3	景区游览	缺失	缺失
天数/天	6	2	娱乐	缺失	缺失
长途交通/元	7764	391.5	购物	缺失	缺失
住宿	6036	699	市内交通	缺失	缺失
餐饮	缺失	缺失	邮电通讯	缺失	缺失

第一步:缺失数据填补

根据上文中所给出的分析,长途交通、市内通讯由于数据之间关联性不强,或者可以说其他数值大小对这两项数据没有影响,因此采用均值插补的方式对这两项数据进行填补。对于此处的实例,由于长途交通数据已经知道,因此只需对市内通讯费用进行均值插补,得到的数值填入表6。

对于餐饮、旅游、购物、娱乐,由于缺失量过多,因此对缺失量中部分采用均值插补,对娱乐项目费用采用神经网络算法进行插补(充分利用已经知道的长途交通和住宿等信息),具体算法流程如上文所述,可以得到两组数据结果如表6所示。

表6.完成补足后数据

类别	实例1	实例2	类别	实例1	实例2
旅行人数/人	8	3	景区游览/元	1789	178.5
天数/天	6	2	娱乐/元	1371	213
长途交通/元	7764	391.5	购物/元	5824	579
住宿/元	6036	699	市内交通/元	1386	253
餐饮/元	3698	441	邮电通讯/元	173	21.6

第二步:聚类层次分析

1.通过对实例1层次分析得到各特征量占据总消费比例:

Basic=(长途交通+住宿+餐饮)/总消费

Purchase=(景区游览+购物+娱乐)/总消费

Phone=邮电通讯/总消费

Short_trans=市内交通/总消费

Basic:0.6240

Purchase:0.3204

Phone:0.0062

Short_trans:0.0494

参照表4给出的聚类中心,计算得到的数据与各个聚类中心的距离。

距离的计算公式如式4.1所示:

$$D = dist(data1, center)^2 + dist(data2, center)^2 + dist(data3, center)^2 + dist(data4, center)^2 + \cdots dist(datan, center)^2 \quad (4.1)$$

将第一类别的聚类中心与得到的各特征量带入式4.1可得到:

$D_1=(0.6240-0.5415)^2+(0.3204-0.3763)^2+(0.0062-0.0074)^2+(0.0494-0.0749)^2=0.01058269$

同D_1的计算方式相同,我们可以得到该数据与其他聚类中心之间的距离:

$D_2=0.00092907$

$D_3=0.00182326$

$D_4=0.01785891$

$D_5=0.11110778$

在所有的找到距离中,找到最小的距离值。此例中D_2值最小,所以该数据属于D_2所示聚类中心,即第二类较高消费旅游者。

由于该项数据是通过后期处理得到的缺失数据,因此可以根据原始数据,进行类别预测。其原始数据和各簇中心的距离为:

$D_1=0.0088$

$D_2=0.0006$

$D_3=0.0033$

$D_4=0.0218$

$D_5=0.1202$

由上述结果可知，D_2 值最小，所以该数据属于 D_2 所示聚类中心，即第二类较高消费旅游者。这与本文算法所得到的结果一致，所以可以证明本文所提出的算法具有较好的缺失数据预测能力。

2.实例 2 聚类分析

各特征量占据总消费比例

Basic：0.5516

Purchase：0.3495

Phone：0.0078

Short_trans：0.0911

同样参照表 3 给出的聚类中心，计算得到的数据与各个聚类中心的距离，得到以下数据：

$D_1=0.00108285$

$D_2=0.00393711$

$D_3=0.01650636$

$D_4=0.04577511$

$D_5=0.1704715$

找到 D 值最小的值，该实例中 D_1 值最小，所以该数据属于 D_1 所示聚类中心，即第一类高消费旅游者。

由于该项数据是通过后期处理得到的缺失数据，因此可以根据原始数据，进行类别预测。其原始数据和各簇中心的距离为：

$D_1=0.0004$

$D_2=0.0055$

$D_3=0.0207$

$D_4=0.0538$

$D_5=0.1862$

由上述结果可知，D_1 值最小，所以该数据属于 D_1 所示聚类中心，即第一类高消费旅游者。这与本文算法所得到的结果一致，所以可以证明本文所提算法具有较好的缺失数据预测能力。

五　研究结论和展望

旅游数据的研究是目前国内外学界研究的热点，而与互联网、大数据相结合的旅游数据研究在国内外都处于起步阶段，有大量的工作可以开展，也有很多的空白需要填补。本次研究建立了一个旅游者消费类型预测模型。该模型对于存在缺失的数据，用回归预测法和插值填补法相结合的方法进行预测，将缺失数据填充之后，对旅游者消费数据进行聚类分析。经过聚类，得到了五种消费者类别，即高消费旅游者、较高消费旅游者、中等消费旅游者、偏低消费旅游者、低消费旅游者。最后，给出了两组数据进行实例分析，通过计算数据到聚类中心的距离判断出这两组数据应属哪个消费层次。最后通过实例分析证明，本文所提模型具有较好的缺失数据预测能力。

本文的理论价值主要体现在对旅游个体消费行为的研究中。现有的旅游消费行为的理论多数起源于营销和消费动机理论，对消费感知、消费态度、消费行为与收入的关系的问题研究较为深入，但对于游客个体的旅游消费行为的预测性研究很少。其原因主要是获取数据较难，而且研究方法没有跳脱出传统统计研究的范畴。本文运用了移动互联网平台采集数据，建立一个预测游客消费行为、消费层次的模型，直接针对旅游者个体，这对于旅游消费行为学理论的完善有很好的补充意义。

在实践价值方面，在大数据的时代背景下，传统数据获取方式受到了强烈的挑战，迫切需要利用新技术新方法对数据进行有效的收集和利用。本文利用 BP 神经网络、均值差补、聚类分析等数学方法，可以在获取旅游者前置消费数据以后预测该旅游者的消费类型。本文采用了以预测为核心的大数据方法，为后来的研究者提供了可供参考的研究思路。

本文仍存在一些需要改进的地方，由于本文提出的模型中所运用的数据全部来自自研平台，导致数据来源比较单一。为增强本文提供方案的外部效度，本文在此处提供一种替代方案。即：由于本文通过实际数据集得到的预测适用于旅游过程中旅客各项消费之间的关联预测，因此当系统获取数据较弱时，可以采用关系向量替代本文中提出的特征向量（各项数据为消费金额占总消费金额的比例），或者设计相似的手工特征算子进行相似的做法。此外，未来工作也将基于关系学习预测进行，希望能够获取更广泛的数据来增加该模型的典型性与代表性。

参考文献：

[1]郝东阳.中国城镇居民消费行为的经验研究[D].长春：吉林大学，2011.

[2]王赛兰，杨振之.面向大数据的旅游微观数据信息平台研究[J].四川师范大学学报(社会科学版)，2015，(1)：54-61.

[3]周文丽,李世平.基于 ELES 模型的城乡居民国内旅游消费结构实证分析[J].旅游科学,2010,(3):29-38.
[4]王媛,黄震方.国内旅游者消费结构及相关行为因素分析——以南京市为例[J].南京师大学报(自然科学版),2005,(4):123-126.
[5]邓清南.四川省国内旅游消费结构探析[J].成都电子机械高等专科学校学报,2005,(2):57-62.
[6]曹新向.河南省国内游客旅游消费变动的分析[J].旅游论坛,2009,(4):583-588.
[7]MENG F, XU Y L, et al. Tourism Shopping Behavior: Planned, Impulsive, or Experiential? [J]. *International Journal of Culture*, 2012,(3):250-265.
[8]PARK K, REISINGER Y, NOH E. Luxury Shopping in Tourism[J]. *International Journal of Tourism Research*, 2009,(2):164-178.
[9]MYUNG-JA K et al. Investigating the Role of Trust and Gender in Online Tourism Shopping in South Korea[J]. *Journal of Hospitality&Tourism Research*, 2013,(3),377-401.
[10]HENDERSON J C, et al. Shopping, Tourism and Retailing in Singapore[J]. *Managing Leisure*, 2011,(16):36-48 .
[11]ZAIDAN E A. Tourism Shopping and New Urban Entertainment: A Case Study of Dubai[J]. *Journal of Vacation Marketing*, 2015,(22) :29-41.
[13]王小彬.基于机器视觉的 SMT 焊点自动光学检测系统研究[D].苏州:苏州大学,2009.
[14]王燕.一种改进的 BP 神经网络手写体数字识别方法[J].计算机工程与科学,2008,(4):50-52.

The Prediction Model of Tourism Consumption Type Based on Tourism Micro-data Platform

WANG Sai-lan[1, 2], YANG Zhen-zhi[a]1

(1. School of Tourism, Sichuan University, Chengdu, Sichuan 610065;
2.The Jincheng institute of Sichuan University, Chengdu, Sichuan 611731, China)

Abstract: This paper builds a tourist consumption type prediction model by applying the abundant data of tourism consumption acquired from tourism micro-data platforms, which is capable of predicting and identifying the types of tourist consumption with partial data absence. Based on the theory of supervised learning, this model first learns from some existing complete consumption data and continually reduces judgment errors through learning algorithm until the model can conduct accurate prediction. According to different data missing, BP neural network and mean value interpolation is applied to replenish them. The resulting data are clustered by K-means clustering and thereby the types and levels of tourist consumption can be predicted. Finally, this model is able to predict the types of tourist consumption even with partial data absence.

Key words: tourism consumption; tourism micro-data platform; tourism comsumption type; prediction model; the theory of supervised learning

[责任编辑:钟秋波]