Hans Journal of Data Mining 数据挖掘, 2015, 5, 1-7
Published Online January 2015 in Hans. http://www.hanspub.org/journal/hjdm
http://dx.doi.org/10.12677/hjdm.2015.51001

Hans汉斯

Kunming Milk Brand Choice Prediction Research Based on Neural Network

Haiyan Li, Yu Fei*

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming
Email: 18388149178@126.com, *feiyukm@aliyun.com, *1350691353@qq.com

Received: Dec. 8th, 2014; accepted: Dec. 30th, 2014; published: Jan. 15th, 2015

Abstract

With the improvement of living conditions, people pay more and more attention to the healthy diet, and their brand consciousness is stronger too. Taking the milk brand for an example, this paper studies the consumers how to make brand choice and discusses the connection between the consumer characteristics and the final choice of milk brand. The result shows that the brand has large influence on consumer behavior. Consumers' gender, age, income, educational level and family structure have strong links with the final brand choice. And the BP neural network model fits the relationship well. Two tests of the model show about 80% of accuracy. The conclusion provides certain reference for milk producers and sellers.

Keywords

Milk Brand, BP Neural Network, Prediction

基于神经网络的昆明市牛奶品牌选择预测研究

李海燕，费　宇*

云南财经大学统计与数学学院，昆明
Email: 18388149178@126.com, *feiyukm@aliyun.com, *1350691353@qq.com

收稿日期：2014年12月8日；录用日期：2014年12月30日；发布日期：2015年1月15日

*通讯作者。

摘　要

随着生活条件的提高，人们越来越注重饮食的健康，品牌意识越来越强。本文以牛奶品牌为例，研究了消费者如何进行品牌选择，讨论了消费者特征与最终选择的牛奶品牌之间的联系，结果显示品牌对于消费者的行为有较大的影响，消费者的性别、年龄、收入、文化程度、家庭结构与最终的品牌选择有很强的联系，可以采用BP神经网络模型来拟合这种关系，模型用于回判分析以及三折交叉检验均达到较好效果，准确率80%左右，为牛奶生产者与营销者提供一定的参考。

关键词

牛奶品牌，BP神经网络，预测

1. 引言

随着我国经济的发展，人们健康意识的提高，食品健康安全越来越成为人们的焦点。牛奶作为一种健康消费品，也越来越得到了人们的关注。从 1980 年到 2013 年我国牛奶产量走势整体呈现上升趋势，特别是从 2000 年到 2007 年经历了快速增长，从 2000 年 827 万吨激增到 2007 年的 3525 万吨，年复合增长率达到 12.78%，2007 年到 2013 年基本平稳，2013 年牛奶产量达到 3531 万吨。据农业部预测，到 2030 年，中国奶类人均占有量将达到 25 千克，总产量达到 4250 万吨，中国将成为世界上最大的牛奶市场。

牛奶已经进入千家万户，成为人们日常生活必备品。消费者是如何选择牛奶品牌，消费者的自身特征是怎样影响消费者的消费行为的，成为了牛奶生产者和营销者迫切关注的问题，本文就牛奶消费者自身特征与最终的牛奶品牌选择之间建立了联系，研究二者之间的关系，建立了预测模型。本文的研究对象是昆明市牛奶消费者，昆明市地处云南，云南具有优越的自然气候条件，形成了独特的牛奶品牌市场，云南除了有蒙牛，伊利，光明等全国知名品牌外，有很多本地品牌，如雪兰、来思尔、蝶泉、欧亚，七彩云等等，并且本地牛奶的知名度在当地居民中很高，鉴于昆明市牛奶市场的独特性，本文研究了昆明市牛奶消费者。文章共分为五部分，第一部分为引言，介绍了我国牛奶业发展状况，以及昆明市独特之处；第二部分为文献综述部分；第三部分介绍了数据来源及其基本情况；第四部分是数据处理及建立模型并进行模型检验部分；最后是结论总结部分。

2. 相关文献综述

在中国乳业大发展的背景下，很多研究者对我国牛奶消费市场及牛奶消费者进行了研究，从研究内容上来看，大致分为三类：

第一类：对牛奶消费市场的目前状况、发展前景以及对策的研究。

近些年我国牛奶市场发展快，牛奶越来越得到人们的认可，再加上我国发生的重大奶产品事件，食品安全也越来越得到人们的关注。很多研究者对我国牛奶消费市场的现状做了分析，并指出我国牛奶消费市场问题还是比较突出的，表现在客观和主观两个方面，客观主要是指牛奶行业本身存在不足，在管理、结构、规模、奶源等方面都存在问题，需要进一步改善，主观方面是指消费者本身的牛奶知识也很缺乏，只是很肤浅的知道牛奶是有营养的，健康的，可以每天饮用，却不知道如何根据自身的健康状况选择符合条件的牛奶。在指出问题的基础上，对比国际牛奶市场，研究者们得出我国的牛奶市场发展前景广阔，表现出对牛奶未来发展的极大信心，并针对问题提除了相应政策建议。例如冯启，张旭(2013) [1]、张洪峰(2011) [2]、刘卫国(2009) [3]、杨志春(2009) [4]、耿莉萍(2012) [5]等学者对我国牛奶市场做了类似

分析。

第二类：对牛奶因素对于最终牛奶消费者的选择的影响的研究

这类研究一般是在研读文献的基础上，选取一些和消费者消费行为有关的牛奶因素作研究，采用一定的方法，将最终的影响因素划分为几大类，或者对所选因素的影响作用进行顺序排列。例如申菊梅(2007) [6]选择了 20 个影响消费者消费习惯的因素，运用主成分分析法，最终找到七大因素，即营养、质量、品牌、消费动机、牛奶粘稠度、价格、以及消费群体，其中营养和质量是影响消费者的两大关键因素；余萍(2013) [7]从营养和安全角度进行了分析。

第三类：对某个地区牛奶消费者特征进行比例或趋势上的统计分析

这类研究是比较多的，都是只对某个地区的牛奶消费者与消费行为相关的一些特征的统计分析，一般选取的特征有消费者的性别，年龄，文化程度，收入，职业类型，购买频率，购买地点，购买目的，购买偏好等等。进行比例统计，能够掌握本地区牛奶消费者的特征分布，以及影响消费者购买牛奶行为的因素所起到的作用，对全面了解牛奶消费者有一定的指导意义。例如姜芳(2007) [8]等人对南京牛奶消费者的研究，胡杨、张哲晰(2013) [9]对吉林市液态牛奶消费的特征分析，张世海(2009) [10]等对北京市的研究。

3. 数据来源及数据基本情况

本次研究所使用数据由小组发放问卷调查得到。在查阅有关文献之后，根据调查目的设计调查问题，并进行抽样调查。本次调查共发放 100 份试卷，其中有效问卷 96 份，有效率为 96%。根据调查结果，被调查者基本特征统计如表 1 所示。

Table 1. Statistical characteristics of respondents
表 1. 被调查者基本统计特征

样本统计特征以及选项		样本数/人	比例
性别	男	39	40.63%
	女	57	59.38%
年龄	15 岁以下	4	4.17%
	15~22 岁	32	33.33%
	23~30 岁	24	25.00%
	31~55 岁	23	23.96%
	56 岁以上	13	13.54%
收入	暂时没有收入	37	38.54%
	2000 元以下	15	15.63%
	2001~5000 元	33	34.38%
	5001~8000 元	6	6.25%
	8000 元以上	5	5.21%
文化程度	初中	13	13.54%
	高中	14	14.58%
	专科	15	15.63%
	本科	45	46.88%
	本科以上	5	5.21%
	其他	4	4.17%
家庭人口数	一人	3	3.13%
	二人	15	15.63%
	三人	44	45.83%
	三人以上	34	35.42%

从统计结果来看，在被调查者中，59.3%是女性，因为大多数家庭女性是日常消费的主导。被调查者的年龄分为五个阶段，分别是 15 岁以下、15~22 岁、23~31 岁、31~55 岁、55 岁以上，其中 15~22 岁、23~31 岁、31~55 岁的比例比较大，合计超过八成，这说明年轻人是主要消费人群，55 岁以上占比 13.54%，这部分消费者是退休人群，在牛奶消费者中也占有一定比例。被调查者收入也分为五个层次，其中没有收入的占据 38.54%，这部分大多数是学生，他们没有收入，除了没有收入的学生以外，收入在 2000~5000 元的是最多的，说明人们的生活水平提高，具有消费潜力。从文化程度上来看，高中和本科的是人数最多的，特别是本科人数占比接近 50%，但是研究生文化程度人数较少，说明我国的本科教育已经大众化。从家庭结构来看，三人及三人以上的家庭是最多的，总计超过 80%。

4. 实证分析

4.1. 方法介绍

人工神经网络(Artificial Neural Networks)是对自然神经网络的一种模拟，可以有效地解决复杂的不明朗的相关变量之间的回归和分类问题。BP(Back Propagation)神经网络是目前应用最广的神经网络模型之一，图 1 是一个三层神经网络示意图，由输入层(input layer)、隐含层(hidden layer)、输出层(input layer)、各层之间的链接权重以及激活函数构成。

一个神经网络模型的建立，关键是链接权重的确定，其求解采用梯度下降法，对最初随机权重不断进行调整，直至收敛。

4.2. 数据处理

4.2.1. 自变量的处理

本次调查所得数据均以选项形式获得，因此所得变量都是分类变量，对于有多种选择的分类变量，可以将其拆分为多个二分类变量，以便于后续工作的开展。按照此原则，以上五个问题的答案，就可以拆分为 17 个二分类变量，具体拆分结果如表 2。

根据上述拆分结果，最终确定的自变量为：性别、年龄 1、年龄 2、年龄 3、年龄 4、收入 1、收入 2、收入 3、收入 4、文化 1、文化 2、文化 3、文化 4、文化 5、家庭人数 1、家庭人数 2、家庭人 3 共 17 个自变量。每个自变量都是一个二分类变量，如此的设定，在本质上，是将每个问题的最后一个选项作为了基础项，其余选项的系数只是和基础选项的影响量化值的差异值，同时有效的避免了自变量之间线性相关。

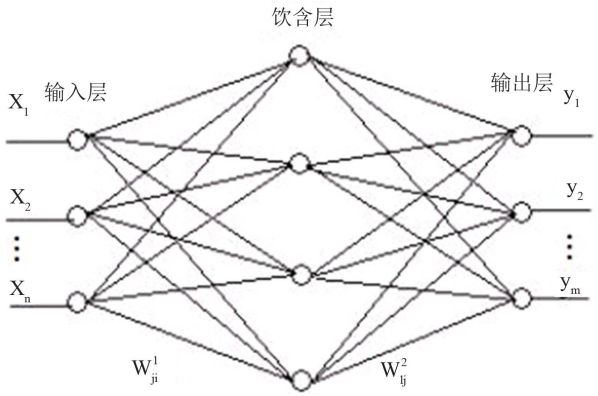

Figure 1. Three layer BP neural network diagram
图 1. 三层 BP 神经网络示意图

Table 2. The independent variables split table
表 2. 自变量拆分示意表

拆分前变量名称及选项	性别		年龄					收入				
	男	女	15 岁以下	15~22 岁	23~30 岁	31~55 岁	56 岁以上	暂时没有收入	2000 元以下	2001~5000 元	5001~8000 元	8000 元以上
拆分后变量名称及表示	性别 =1	性别 =0	年龄 1 =1	年龄 2 =1	年龄 3 =1	年龄 4 =1	年龄 1 = 0 年龄 2 = 0 年龄 3 = 0 年龄 4 = 0	收入 1 =1	收入 2 =1	收入 3 =1	收入 4 =1	收入 1 = 0 收入 2 = 0 收入 3 = 0 收入 4 = 0
	性别		年龄 1	年龄 2	年龄 3	年龄 4		收入 1	收入 2	收入 3	收入 4	

拆分前变量名称及选项	文化程度						家庭人口数			
	初中	高中	专科	本科	本科以上	其他	一人	二人	三人	三人以上
拆分后变量名称及表示	文化 1 =1	文化 2 =1	文化 3 =1	文化 4 =1	文化 5 =1	文化 1 = 0 文化 2 = 0 文化 3 = 0 文化 4 = 0 文化 5 = 0	家庭人数 1 =1	家庭人数 2 =1	家庭人数 3 =1	家庭人数 1 = 0 家庭人数 2 = 0 家庭人数 3 = 0
	文化 1	文化 2	文化 3	文化 4	文化 5		家庭人数 1	家庭人数 2	家庭人数 3	

4.2.2. 应变量处理

根据各牛奶品牌在昆明市牛奶市场占有率的情况，本次研究将蒙牛，伊利，雪兰各作为一种品牌，将其余牛奶品牌合并为一类，称为其他，具体结果表示如表 3。

表中可以看出在云南地区，雪兰的占有率是有大的，其次为蒙牛，光明作为全国性品牌表现力相对较差，来思尔、欧亚、七彩云、蝶泉均为云南本地牛奶，市场占有率对比雪兰差距很大。进口牛奶在云南当地的市场占有率是很小的。

4.3. 模型建立

本次研究目的是建立消费者特征与最终选择牛奶品牌之间的联系，自变量为上述 17 个二分类变量，应变量为 4 种牛奶品牌，采用 BP 神经网络算法建立模型。

BP 神经网络隐含层节点数的确定[11]，常用方法为试凑法，根据经验公式 $m=\sqrt{n+l}+a$ 进行试验，式中 m 代表隐含层节点数，n 和 l 分别代表输入层节点数和输出层节点数，本次研究分别为 17 和 4，a 是一个常数，介于 1~10 之间，根据此公式，计算隐含层节点数为 5~14 个，逐一进行多次试验，得到最佳隐含层节点数为 8，最终确定此模型是一个 17-8-4 的 BP 神经网络模型。

利用已建模型，对样本进行回判分析得到混淆矩阵如表 4 表示。

混淆矩阵的行指标表示实际值，列指标表示预测值，例如矩阵第一行第二列的数以 1，表示实际最终选择是“其他”牛奶品牌，而预测结果是“蒙牛”品牌。据此，只有矩阵对角线位置表示预测结果和实际结果是一致的，非对角线位置表示预测是错误的。故预测错误率为 0.208333 预测正确率为 0.791667。回判分析结果表明，BP 神经网络模型能够较好的反应消费者自身特征与最终牛奶品牌选择之间的联系，预测错误率约为 21%。

4.4. 模型检验

模型检验采用三折交叉检验法，三折交叉检验大致分为两步[12]：

1) 将样本分为三个子样本。

将所有的样本，按照应变量，先进行分类，然后将每一类的样本分为三份，最后在每一类中拿出一份放在一起，形成一个子样本，根据这种方法，最终可以得到三个类似的子样本，该类子样本在理论上来说，选择各牛奶品牌的人数比例和原样本是相等的。

2) 测试

在三个子样本中，轮流抽出一个子样本作为测试集，其余两个子样本作为训练集，比较训练集和测试集的预测错误率，最终结果如表 5 所示。

交叉验证结果显示，训练集和测试集的错误率相差较小，都控制在 15%~31%内，在 20%左右较集中，与回判错误率相差较小，说明模型较稳定，能够较准确地表示消费者自身特征与最终牛奶品牌选择之间的联系，可以作为预测模型对潜在消费者会选择那种牛奶品牌作出判断。

5. 结论

本文将 BP 神经网络模型引入到牛奶消费者品牌选择研究中，分析了牛奶消费者自身特征性别、年龄、文化程度、收入、家庭常住人口数与消费者最终选择的牛奶品牌之间的关系，在做变量处理时，本文采用了将所有自变量化为二分类变量，最终得到 17 个自变量，应变量做处理时，将市场占有率小的牛奶品牌合并为一类牛奶品牌，形成四大品牌，建立二者之间的预测模型，为牛奶生产者和销售者提供相关建议。

本次研究得到的 BP 神经网络模型，能够较稳定的对已知消费者特征的牛奶消费者作出牛奶品牌选择预测，这说明，牛奶消费者的本身特征与最终的牛奶品牌选择之间是存在强联系的，因此对牛奶生产者和销售者给予以下建议：

1) 作为牛奶生产者，在进行牛奶品牌建立时，应先确定目标人群，分析目标人群的特征，根据目标

Table 3. Response variable process table
表 3. 应变量处理示意表

合并后品牌	蒙牛	伊利	雪兰	其他					
合并前品牌	蒙牛	伊利	雪兰	光明	来思尔	欧亚	七彩云	蝶泉	进口牛奶
比例(%)	23.96	19.79	34.38	21.87					

Table 4. Confusion matrix
表 4. 混淆矩阵

	蒙牛	其他	雪兰	伊利
蒙牛	18	1	2	2
其他	5	14	2	0
雪兰	3	0	29	1
伊利	1	3	0	15

Table 5. Three fold cross test results
表 5. 三折交叉检验结果

	第一次错误率	第二次错误率	第三次错误率	平均值
训练集	0.2343	0.1563	0.1875	0.1927
测试集	0.1563	0.3125	0.2500	0.2396

人群的需要生产商品，是成功的关键。

2) 作为牛奶销售者，在选择牛奶品牌之前，应确定牛奶销售对象是什么人群，例如建立在学校内的商店，和小区内的商店，二者的服务对象是不一样的，年轻人会更喜欢伊利，蒙牛等国内知名品牌，而小区内可能会更喜欢本地牛奶品牌。针对大型超市，其服务对象覆盖全面，就应该在牛奶品牌摆设，以及牛奶存储等方面下工夫，根据超市内每天进入的消费者特征来确定摆设方式和存储量，达到既能满足消费要求，又不至于某种牛奶长久囤积，增加收益，减小成本。

基金项目

云南省省院省校教育合作咨询共建重点学科——统计学(42111217003)。

参考文献 (References)

[1] 张旭, 冯启 (2013) 中国乳企的战略布局与发展思路分析. *乳品与人类*, **1**, 5-12.

[2] 张洪峰 (2011) 液态牛奶市场消费行为及市场前景浅析. *产业经济*, **7**, 90-91.

[3] 刘卫国 (2009) 我国液态奶市场现状、问题及发展对策. *企业家天地下半月刊*(*理论版*), **6**, 41-42.

[4] 杨志春 (2009) 牛奶消费的思考. *行业论坛*, **12**, 73-74.

[5] 耿莉萍 (2012) 我国乳品质量安全问题频发的原因与对策. *北京工商大学学报*(*自然科学版*), **1**, 74-80.

[6] 申菊梅, 李欣 (2007) 某市牛奶消费者消费习惯分析. *科技资讯*, **11**, 219-220.

[7] 余萍, 范志红, 李萌, 龙菲平 (2013) 营养和安全因素对北京消费者牛奶产品选购的影响. *乳业科学与技术*, **6**, 14-17.

[8] 姜芳, 吴阳裕, 王小虎, 丁月云, 张莉莉, 王恬 (2007) 关于南京市牛奶消费的调查报告. *畜牧与兽医*, **4**, 27-29.

[9] 胡杨、张哲晰 (2013) 吉林市居民液态牛奶消费者的特征分析. *商业研究*, **19**, 120-121.

[10] 张世海, 王雅春, 俞英, 潘敦菲, 郭黎宁, 和思淼, 王鹏, 齐思锦, 牛牧田 (2009) 北京市家庭奶制品消费调查. *中国乳业*, **11**, 36-37.

[11] 周瑛, 刘天娇 (2013) 基于神经网络的高校图书馆知识服务评价体系研究. *情报理论与实践*, **2**, 55-59.

[12] 吴喜之 (2012) 复杂数据统计方法－基于 R 的应用. 中国人民大学出版社, 北京, 41-43.

汉斯出版社为全球科研工作者搭建开放的网络学术中文交流平台。自2011年创办以来，汉斯一直保持着稳健快速发展。随着国内外知名高校学者的陆续加入，汉斯电子期刊已被450多所大中华地区高校图书馆的电子资源采用，并被中国知网全文收录，被学术界广为认同。

汉斯出版社是国内开源（Open Access）电子期刊模式的先行者，其创办的所有期刊全部开放阅读，即读者可以通过互联网免费获取期刊内容，在非商业性使用的前提下，读者不支付任何费用就可引用、复制、传播期刊的部分或全部内容。