

**南开大学第二届“校长杯”创新创业大赛数学-统计-人工智能专项赛学术论文**

论文题目： 基于结合算法对分布不均小样本数据的分析

团队名称： 淇奥团队

负责人： 王子旭

学院： 软件学院

参赛类别： 数学-统计-人工智能专项赛

材料上交日期：2020年4月30日

**摘要**

为改善小样本数据统计分析得到的结果精度较低的问题，开展利用机器学习基础方法对样本数据抽象分析的研究。属性分布不均的小样本数据往往不能反映整体问题，也无法明确判断对整体数据而言影响最大的因素，也不容易抽象出来特征和标签用以满足监督学习。利用机器学习中的聚类算法，并结合监督学习对数据的各种特征进行权重计算与比较，找到其中影响最大的核心特征。应用**K-means**算法确定集群数目和聚集中心点，并将其作为**logistic**回归的标签，既利用监督学习的优良模型提高了聚类算法的准确度，又在聚类算法的帮助下为监督学习确定了合适的标签。针对影响大学生选购智能手机的调查问卷数据，进行统计图与算法结果的对比分析，完成初步对这类结合算法可行性的验证。

关键词：小样本数据，机器学习基础，聚类算法，**logistic**回归，算法可行性

**Abstract**

In order to solve the problem of low accuracy when pushing statistical analysis to a small size data set, we decide to carry out the research on using basic machine learning method to abstractly analyze the data with small sample size. Genuinely, a data set with small sample size and unbalanced-distribution attributions can neither reveal some global laws nor find the most important factors, let alone extract features and labels for later supervised learning. By taking advantages of clustering algorithm in machine learning and combining it with supervised learning, we can find the most crucial features according to the value of features’ major which is an attribution set to notify the importance of all the features. After executing K-means algorithm we can get the value of K and the centroids. Then we number the clusters and label it to the samples. Next, we can use the features and labels above to implement one-Vs-all logistic regression. This way, we combine the advantages of both supervised and unsupervised learning algorithm. Lastly, based on processing the data that may affect university students’ smart phone purchasing choice we got from questionnaire, we compare our machine learning algorithm’s result and tradition statistical analysis to prove the validity of our combine algorithm preliminary.

Key words: data with small sample size, machine learning foundation, clustering algorithm, logistic regression, validity of the algorithm

**目录**

**第一章 绪论**

1. 机器学习算法的发展历史及现状
2. 论文的研究目的、意义和主要内容
3. 研究目的和意义
4. 主要研究内容

**第二章 结合算法的基础理论**

1. 引言
2. 结合算法的设计和实现

**第三章 问卷数据直观分析和统计图**

1. 性别影响因素
2. 大学生对智能手机的品牌忠诚度
3. 提价不同额度的敏感程度
4. 国产手机与苹果手机价格范围比较
5. 大学生月均消费和手机价位之间的关系

**第四章 结合算法处理数据**

1. 非监督学习
2. 监督学习
3. 分类器结果对比
4. 性别影响因素
5. 智能手机品牌忠诚度

**第五章 结合算法的评价与改进**

1. 结合算法的优劣
2. 影响结合算法可信度的因素
3. 结合算法的改进与展望

**第一章 绪论**

本章首先对机器学习基础算法的发展历史进行了回顾，分析了算法改良的现状，介绍了数据的直观分析方法，然后提出了本论文的研究目的和意义，最后介绍了主要研究内容。

1. **机器学习算法的发展历史及现状**

机器学习是人工智能研究的核心内容，它的应用已遍及人工智能的各个分支，如专家系统、自动推理、自然语言理解、模式识别、计算机视觉、智能机器人等领域。而机器学习的发展又分为如下几个阶段:知识推理期、知识工程期、浅层学习和深度学习。

知识推理期起开始于 20 世纪 50 年代中期，此时机器学习主要通过专家系统赋予计算机逻辑推理能力，和实现的自动定理证明系统证明了逻辑学家和的《数学原理》中的其中 52 条定理。20 世纪 70 年代开始，机器学习进入知识工程期， 作为知识工程之父在 1994 年获得了图灵奖，由于人类的总结能力有限，因此机器学习陷入了关于获取新知识的瓶颈期。

其实，从20 世纪 50 年代开始，就已经有机器学习有关“神经网络”的相关研究，代表性工作主要是基于神经感知科学提出的计算机神经网络，且这项工作引起了随后十年机器学习的火爆。但是由于计算机的运算能力有限，通常都是只有一层隐含层的浅层模型，并且随着KNN等相继提出，浅层模型在模型理解、准确率、模型训练等方面被超越，而机器学习的发展几乎处于停滞状态。

2006 年，机器学习正式进入了深层网络实践阶段。同时，云计算和 GPU 并行计算也为深度学习的发展提供了基础保障，近年来在各个领域都取得了很大的发展。[1]

本论文中使用的是K-means算法和logistic回归，均属于机器学习初期，即知识推理期被提出的重要理论方法，有足够充分的理论支持和强大的运算能力及精度。“K-means”最早是Hugo Steinhaus在1957年提出的，James MacQueen在1967年正式提出。标准算法最初由 Stuart Lloyd 在1957年提出，但直到1982年才在贝尔实验室以外的地方公开发表。[2]

在统计学中，logistic模型是一个回归模型，根据分类的个数不同有二元回归和多元回归之分。逻辑回归是由统计学家David Cox在1958年提出的，我们实验中使用的二元逻辑回归模型常被用来估计基于多个独立变量的二元反应的概率。

逻辑回归可以看作是广义线性模型的一个特例，因此它与线性回归很像。但是逻辑回归模型是基于因变量和自变量之间的关系这个假设而运行的，与线性回归的完全不同。

逻辑回归作为一种广义的线性回归模型, 具有诸多理想的统计特性和充 分的可解释性, 易于理解和实现并且很适用于自变量具有连续性和类别性的情况，正因如此，目前应用于多个领域，包括机器学习、医学、社会学等等。[3]

拉瑞·瓦瑟曼（）认为“机器学习和数理统计关心的是同一件事，即我们能从数据中学到什么”，并指出统计学中的评估、分类器、数据点、回归与分类、协变量、响应理念可以分别对应于机器学习当中的学习、假设、用例、有监督学习、特征、标记理念。[4]

机器学习与传统计量经济学方法的结合是目前经济学应用机器学习进行预测的一个明显趋势。[5]机器学习算法凭借其强大的样本之外的预测能力、对异构数据和大规模数据的处理能力，以及对数据质量的低敏感度的处理方法，在经济研究中具有“更好地让数据说话”的功能。

**（二）论文的研究目的、意义和主要内容**

1.研究目的和意义

由联系并不紧密的特征组成的现实数据，很难直接抽象出机器数据，难以对数据进行有效分析，即使能够得到一定结论也会因没有固定模型而降低准确率。通过结合监督学习和非监督学习，将二者的优势发挥出来，最大可能完成由少量样本、特征量不强带来的整体数据分析上的困难。

本次研究因为建立在选取特征均为显著影响因素，且样本数据可靠性能够保障，因此通过多次调参得到较为稳定的最优聚类方式，即全局最小值点。

对本论文的具体研究课题来说，利用简单信息进行对大学生影响手机购买因素的研究，从中筛选出核心影响因素，使智能手机的生产和销售在营销政策上的针对性有所提高，有助于智能手机品牌定位化和目标化。同时筛选出来的主要因素可作为用户个性化定制的重要参考指标。

2.主要研究内容

智能手机已成为互联网时代必不可少的工具。中国目前已经完成了智能手机的普及，国内市场也已接近饱和。智能手机在大学生中有着极大的市场。但手机品牌多种多样，不同的品牌性能不同、定位不一，针对的主要购买人群也有所区别。

尽管品牌逐渐多样化，但人们对其品牌忠诚度却大有差别。市场调研机构IDC发布的数据显示，2018年苹果和三星的在国际市场份额之和达到了35.7%，并且掌握着最具价值的中高端市场。由于其产品质量、操作系统、研发实力等过人之处使消费者对其产生了较高的品牌忠诚，产生了极强的市场溢价能力。[6]作为“国货之光”的华为，在国内智能手机市场也占有较大的市场份额。

对智能手机来说，品牌不同、型号不同、功能有差距，运行速度和使用感受都会导致价格上出现分层。通过对大学生的月均消费水平和所使用手机情况的调查，可以从某种程度上反映出部分大学生对智能手机的消费观念。

设计一份有关月均消费水平和手机使用情况的问卷，内容包括“性别”、“年级”、“月均消费”、“手机品牌”、“手机价位”、“提价后的购买意愿”以及“是否为第一部智能手机”。通过回收的160份有效问卷数据，希望能够找到影响大学生手机选择的主要影响因素。

由于数据分布没有明显规律，各变量之间无必然联系，还可能存在特征冗余的问题。通过相关变量组合统计分析，可以看出一些规律，研究由变量组合产生的子问题，但不能对整体数据有较好的全局观，并且在面对大量需要处理的数据时，直观分析会耗费大量时间和精力。通过机器学习训练，可以较为简单且快速有效的找到影响大学生选择手机的部分重要因素。设计满足条件的机器学习算法，将利用该算法的学习成果和直观统计分析结果对比，结论一致，初步证明算法有效。

**第二章 结合算法的基础理论**

本章介绍了机器学习中两种基本方法，即非监督学习和监督学习，分析了结合算法将这两种学习过程结合在一起的设计思路。

1. 引言

大数据时代的到来使得机器学习领域焕发勃勃生机。但面对小样本数据，人们会更倾向于用统计学分析数据，对较少的特征维度进行研究。多维特征数据就会被分为几个子问题而进行多角度研究分析，最后得到某些符合实际情况的结论。利用机器学习对数据的处理分析能力，就可以从多个特征量中择出影响最大几个。聚类算法可以对全体有效数据初步处理，大致分类，为后续的模型的优化训练提供标签。再利用logistic回归分析或多层神经网络技术，找到最能体现数据样本规律的部分属性。针对属性模糊、分布不均以及数据量较小的问题，使用机器学习这两种基本方法的结合算法，不仅可以看出整体数据的潜在规律，不同分类器的对比也可以实现多角度组合特征，完成子问题的分析验证。

1. 结合算法的设计和实现

首先需要对数据预处理，将定类数据进行虚拟变量处理，也叫哑变量化，也就是将质变量化，通常取值为“0”或“1”。对于有n种选择的某个因素，当将虚拟变量引入模型时，需要设置n-1个虚拟变量，以避免完全的多重共线性。引入虚拟变量可使模型变得更复杂，但使问题描述更加简明，而且接近现实。对于有序变量，如果样本量不够大时，哑变量化会造成变量数目上升，使回归结果变得不可靠。

然后是对整体数据进行聚类，观察集群数与代价函数之间的关系，为使研究问题分析过程简单化，要确定合适的集群个数，同时要兼顾代价函数在可控范围内，以保证最终结果的准确性。

确定分类器个数后，也就是确定了适当的集群数目k，将数据按照某一特定顺序排列，以中间组数据为中心，按照高斯分布,其概率密度函数为：

公式(2.1)

随机选择k个聚类中心。使用K-means算法完成聚类过程。给定n个点和集合，定义目标函数

公式(2.2)

其中表示点到集合的距离，K-means问题使该目标函数最小，即

公式(2.3) [2]

完成聚类后，确定了最终的k个聚类集群，并按顺序标记，将分类号作为监督学习的标签y。

此时经过二次处理的数据已经满足了监督学习所需要的特征和标签。采用多元logistic回归模型，循环进行多次二元分类来完成对数据的多元分类处理。我们以循环中的第t次二元分类为例:特征数为n，样本数为m，聚类算法中的类别数为K，在logistic回归中样本特征的参数向量称为 ；某个样本的特征组成的特征向量称为) ;由样本向量组成的设计矩阵为；样本在第t轮二元分类中对应的标签向量：；其假设函数为：

公式(2.4)

代价（损失）函数为：

公式(2.5)

代价函数的梯度为：

公式(2.6)

得到代价函数和其梯度后，即可以使用梯度下降法或者高级的函数来进行最优化处理，第t类的最佳的线性分类边界参数；更新标签y来继续训练新的二元分类器，最终会得到所有的参数向量。每个以行向量的形式放入矩阵中，即可得到一个K \* (n+1)规模的K元分类器的边界参数矩阵，即

;

样本按照6:2:2的比例分为训练集、交叉验证集和测试集。其中训练集样本称为其样本量为,对应的标签为；交叉验证集的样本称为样本量为，对应的标签为；测试集的样本称为其样本量为，对应的标签为；首先使用所有的来构成设计矩阵，根据它进行上述K元分类器的训练，得到参数矩阵； 接着使用构成的矩阵进行预测，预测得到的结果为 = \*，其中（i，j）表示样本i属于第j类的概率，即

( i , j ) = 公式(2.7)

不难看出样本i的分类结果即为

= 公式(2.8)

于是我们得到了对交叉验证集的样本的分类预测向量predict，将其同进行比较即可得到由训练集得到的分类器的准确率。同理使用测试集数据进行同样的操作，进一步分析准确率。

在经过训练集、交叉验证集和测试集的预测和检验后，得到参数矩阵。由于不同特征对标签的影响程度不同，每一个特征所对应的参数大小也不同。参数越大，对应的特征量对标签的影响就越大。为了保证分类结构的可靠性，进行多次训练，每次训练后总是在结果中选择参数最大的三个特征。设置变量major记录每个特征在每次训练中的参数是否位于前三位，逐次叠加，全部训练结束后，评估分类器的正确率，选出major前三位所对应的特征。最终认为这三种特征量是影响整体数据的核心因素。

**第三章 问卷数据直观分析和统计图**

本章具体对问卷所得的有效数据进行了统计和分析，利用饼状图、箱型图、雷达图和簇状图描述了性别影响因素、大学生对国产手机和苹果手机的品牌忠诚度以及月均消费与使用手机之间的关系。

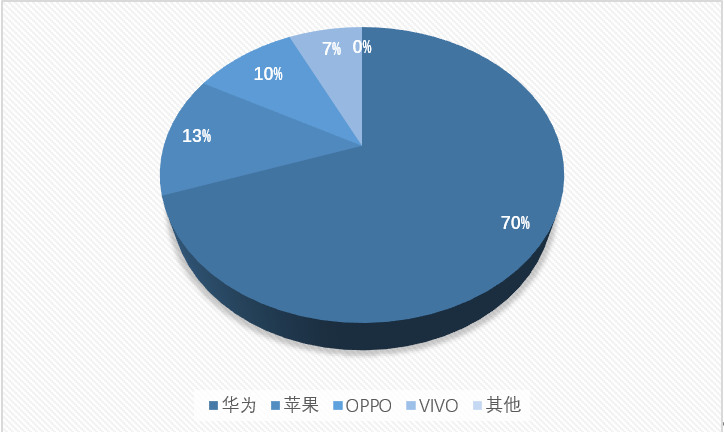
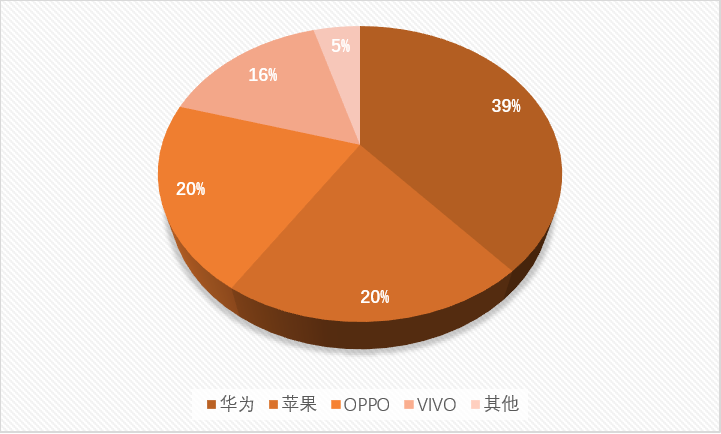
### 表 (3.1) 部分样本数据

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 性别 | 年级 | 月均消费 | 手机品牌 | 手机价位 | 提价是否购买 | | 是否是第一部 |
| 500元 | 1000元 |
| 女 | 本科二 | 507 | 苹果 | 6902 | 是 | 是 | 是 |
| 男 | 本科二 | 303 | OPPO | 2783 | 是 | 否 | 否 |
| 女 | 本科二 | 2040 | OPPO | 1053 | 否 | 否 | 否 |
| 女 | 本科四 | 1782 | 华为 | 2980 | 是 | 否 | 否 |
| 男 | 本科二 | 1123 | 华为 | 7148 | 是 | 否 | 否 |
| 女 | 本科二 | 1974 | 华为 | 2699 | 是 | 是 | 否 |
| 男 | 本科二 | 475 | 华为 | 1815 | 是 | 否 | 否 |
| 女 | 本科二 | 1051 | 其他 | 4448 | 是 | 是 | 否 |
| 女 | 本科二 | 1074 | 华为 | 2334 | 是 | 否 | 否 |
| 女 | 本科二 | 2101 | 苹果 | 7708 | 是 | 是 | 否 |

1. 性别影响因素

按照性别筛选各种手机品牌使用人数的占比。

### 图 (3.1) 女生手机品牌占比 图 (3.2) 男生手机品牌占比



通过图3.1和图3.2的对比，可以看出，女生相比男生来说，对手机品牌的偏好程度较均匀，男生有70%的人数都使用华为手机。综合来看，华为在大学生中市场份额相当大，具有良好的发展潜力和较大的认可度。

1. 大学生对智能手机的品牌忠诚度
2. 提价不同额度的敏感程度。

利用箱型图分析大学生对提价500和提价1000的价格敏感度。

### 图 (3.3) 大学生对各品牌手机的价格敏感度

价格敏感度（price- sensitive)表示为顾客需求弹性函数，即由于价格变动引起的产品[需求量](https://baike.baidu.com/item/%E9%9C%80%E6%B1%82%E9%87%8F/2461773" \t "/Users/prince/Documents\x/_blank)的变化。从图中可以看出提价1000元相比500元的价格敏感度明显提升。其中，苹果手机变化程度是最小的，这也印证了苹果手机相比国产手机品牌具有一定数量的品牌忠诚者。

2.国产手机与苹果手机价格范围比较

将数据按手机品牌进行筛选得到雷达图，用以描述各品牌之间的价格竞争力。

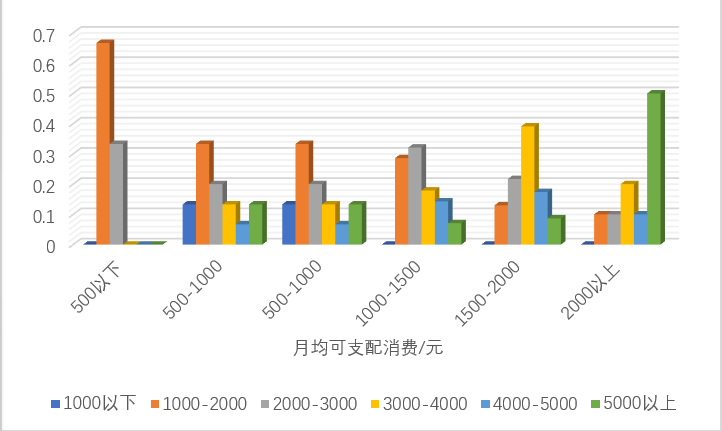
### 图 (3.4) 各品牌手机的价格范围比较

由雷达图易看出苹果手机半数以上的价格都在5000元以上，国产手机相比苹果手机差价较大，由此体现出苹果手机的中高端定位以及良好的溢价能力。

1. 大学生月均消费和手机价位之间的关系

按照月均消费筛选数据，找到月均消费与手机价位之间的相关关系

**图(3.5)大学生月均可支配收入和手机价位的关系**



从图(3.5)中可看出，大学生月均消费与手机价位之间没有必然联系，但会在一定程度上均受消费观念影响。月均消费2000元以上的人中有近半数人使用的手机价位在5000元以上；而月均消费水平较低的人群使用的手机价位多在中等及以下水平。

**第四章 结合算法处理数据**

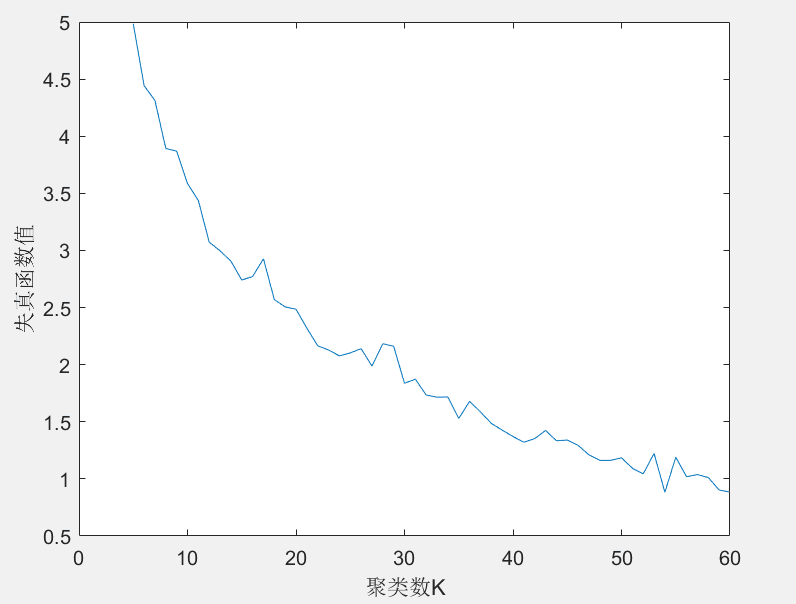
本章将结合算法应用到具体实例中，展现了完整的数据处理过程，并将算法所得结果和前一章统计结论做比较，初步验证了算法的可靠性。

1. 非监督学习

首先将定类数据进行哑变量化处理。变量“性别”的取值分别为男和女，构造男性为“1”、女性为“0”。对“年级”、“手机品牌”、“提价后的购买意愿”和“是否为第一部智能手机”这些变量均照此构造虚拟变量。

对全部数据进行聚类，得到集群数目和失真函数之间的关系。

**图（4.1） 集群数目-失真函数**



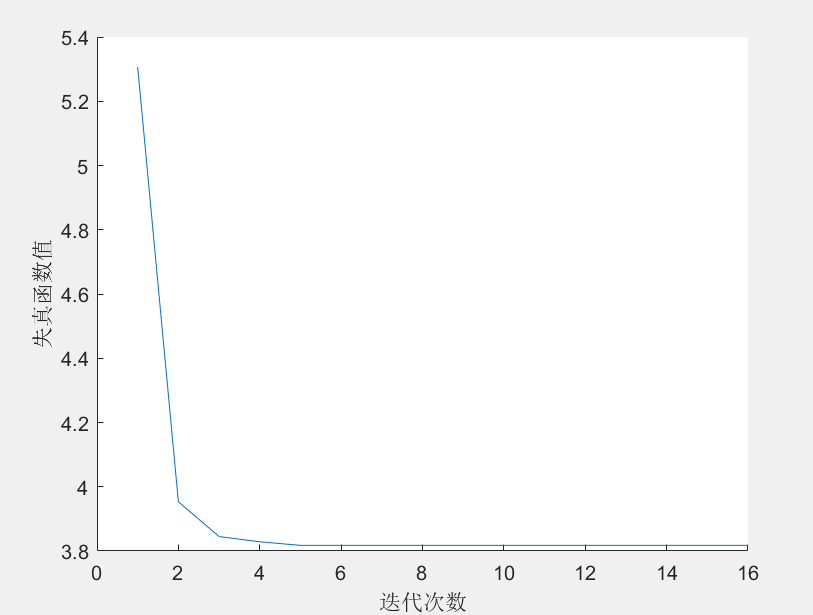
为使后续对数据的分析更加简单，集群数目定为10个。对全部数据进行字典序排列，以第80个数据为中心，通过高斯分布随机选取10个初始聚集中心点,如表(4.1)。

### 表（4.1）量化的部分样本数据

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 性别 | 年级 | 月均消费 | 手机品牌 | 手机价位 | 提价是否购买 | | 是否是第一部 |
| 500元 | 1000元 |
| 1 | 2 | 1918 | 20 | 3201 | 1 | 1 | 0 |
| 1 | 2 | 1510 | 20 | 4410 | 1 | 1 | 1 |
| 0 | 2 | 1431 | 20 | 2499 | 1 | 0 | 0 |
| 0 | 4 | 1988 | 30 | 8795 | 1 | 1 | 0 |
| 0 | 2 | 1950 | 20 | 3385 | 1 | 1 | 0 |
| 0 | 2 | 1144 | 30 | 9128 | 1 | 1 | 0 |
| 0 | 2 | 2049 | 30 | 8283 | 0 | 0 | 0 |
| 1 | 1 | 2000 | 20 | 8900 | 1 | 1 | 0 |
| 1 | 2 | 1455 | 20 | 3696 | 1 | 0 | 0 |
| 0 | 3 | 1915 | 20 | 4318 | 1 | 1 | 1 |

由于不同数据之间差别较大，需要对数据进行归一化处理。完成迭代后，画出迭代次数-失真函数图像。

### 图（4.2）迭代次数-失真函数



运行结束后，“去归一化”得到最终的10个聚集中心点。以编号①~⑩作为其编号。

### 表 (4.2) 归一化处理后的部分样本数据

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **X1** | **X2** | **X3** | **X4** | **X5** | **X6** | **X7** | **X8** |
| 1.0000 | 1.9500 | 1.5709e+03 | 18.9500 | 3.1864e+03 | 0.9500 | 1.0000 | 5.5511e-17 |
| 0.9583 | 2.1667 | 1.2515e+03 | 15.4167 | 2.6378e+03 | 0.7500 | 0.5417 | 1.0000 |
| -2.7756e-16 | 2.5135 | 1.2683e+03 | 16.1351 | 2.3172e+03 | 0.6216 | -1.1102e-16 | 0.0811 |
| 0 | 2.0000 | 2049 | 30.0000 | 8283 | 0 | 0 | 0 |
| 1.0000 | 1.4444 | 1.5078e+03 | 23.3333 | 7.7046e+03 | 1.0000 | 0.4444 | 0 |
| -5.5511e-17 | 1.5714 | 1.2906e+03 | 27.1429 | 7063 | 1.0000 | 1.0000 | 0.1429 |
| 0 | 2.0000 | 1.4833e+03 | 17.4211 | 2.9265e+03 | 1.0000 | 1.0000 | 5.5511e-17 |
| 0.4286 | 3.1429 | 2.2133e+03 | 30.0000 | 8.0366e+03 | 1.0000 | 1.0000 | 0 |
| 1.0000 | 2.0000 | 1.2006e+03 | 19.7143 | 2.8673e+03 | 0.4762 | -2.2204e-16 | 5.5511e-17 |
| 1.1102e-16 | 2.1333 | 1.5448e+03 | 16.4667 | 3367 | 1.0000 | 0.7333 | 1.0000 |

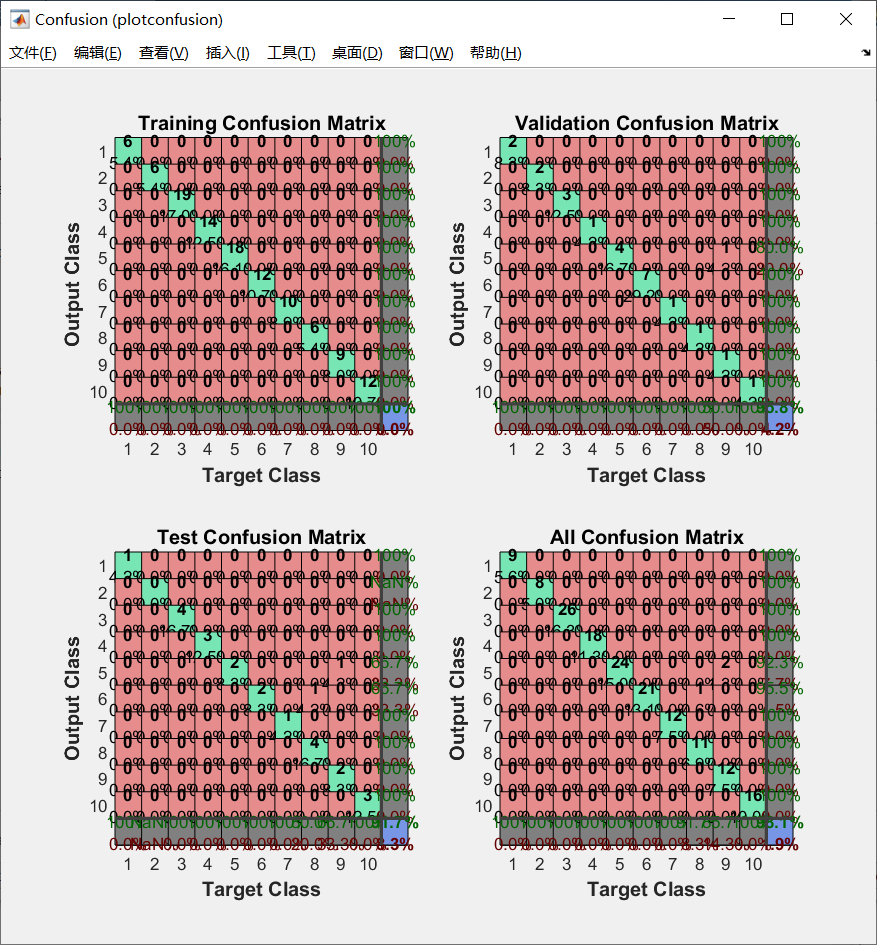
1. 监督学习

用logistic回归进行一对多监督学习，得到的参数。根据设置记录器major，用来记录多次训练后某个特征量对应参数值位于前三的次数，选出major较大的前三个数值。这就最大程度保证了对构造的分类器影响最大的特征量的准确率与可信度。

运行过程需要利用MATLAB神经网络工具箱进行参数调整。对神经元的个数选择需要人工调参，最后调参结果，Test Set错误率仅为8.3%。神经元个数调参过程：

1->30->50->100->40->60->45->50->55->47->50->53->49->50->51->50

### 图（4.3）MATLAB调参结果显示



经过10次训练后，各特征量对应得到的major如表（4.3）。

### 表（4.3）各特征对应的major指数

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **X1** | **X2** | **X3** | **X4** | **X5** | **X6** | **X7** | **X8** |
| 10 | 3 | 2 | 4 | 5 | 3 | 6 | 7 |

从表中易得变量X1、X7、X8对整体数据分类影响最大。这三个变量分别对应调查问卷中的“性别”、“提价1000元后的购买意愿”、“是否为第一部智能手机”这三个量。“是第一部智能手机”评价了大学生在选购新手机时对某一手机品牌质量的信任程度。

由此可得出结论，大学生选购智能手机的主要影响因素为性别、高价格敏感度和手机品牌信赖程度。

1. 分类器结果对比
2. 性别影响因素

对于①号和③号两个聚集集群，在性别和提价1000元的价格敏感度上被明显区分开来。

**表（4.4）**①号③号聚集群的比较

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | 性别 | 聚集数量 | 提价1000元购买意愿 | 是否为第一部智能手机 |
| ① | 男 | 21 | 是 | 不是 |
| ③ | 女 | 37 | 否 | 不是 |

从表（4.4）中可看出女生相比男生在提价1000元时要更加敏感。因此在面对同样的价格时，女生可能会考虑更多的手机品牌。这和统计图中女生的手机品牌比例较均匀相对。

2.智能手机品牌忠诚度

分类器⑦号和⑧号，显示出了苹果手机和国产手机在价格上竞争力的不同，涉及的月均消费水平也从一定程度影响了智能手机的消费观念，但整体影响不大。

**表（4.5）**⑦号⑧号聚集群的比较

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 编号 | 聚集数量 | 月均消费水平 | 手机品牌 | 手机价位 | 提价1000购买意愿 |
| ⑦ | 19 | 2000以下 | 国产 | 5000以下 | 否 |
| ⑧ | 7 | 2000以上 | 苹果 | 5000以上 | 是 |

国产手机大多数在5000元以下，苹果手机恰恰相反，5000元以下很少。但面对提价1000元的情况，苹果手机持有者仍旧愿意坚持苹果品牌；而国产手机在较高价格水平上明显不敌苹果手机。

**第五章 结合算法的评价与改进**

本章结合实例分析了结合算法在应用过程的优点与缺点，以及影响其准确度的因素，最后对结合算法中高级算法的应用改进和实践进行展望。

（一）结合算法的优劣

相比直观分析和统计图法，结合算法可以对全部数据进行筛选和总结，通过对“大学生选购手机影响因素”的数据实践，算法的最终结果可以从一系列特征量中选出影响较大的几个；并且不同分类器之间的比较也可以满足由统计图中得到的结论。合与分都是可估量的，满足了多角度对数据的分析的要求。

从算法本身角度来看，无监督学习中，聚类算法是将n个对象分成k个簇，每个簇中具有相似度较高的部分对象，由此出发进行相关数据分析。在现有聚类算法中，K-means算法实现简单，普及性强，但是聚类集群数目需要提前确定，而且聚类中心点的随机选取也会影响聚类结果，产生局部最优解，影响聚类正确性。[7] 监督学习则可以根据已有的数据，确定输入和输出后，训练得到一个较优的模型。监督学习中训练数据要求既有特征又有标签，这样之后在面对只有特征没有标签的数据时，可以判断出标签。但对于类似问卷来说，标签量难以确定。二者结合，既在一定程度上弥补了无监督学习的准确率不高的问题，又可以给定明确的标签进行监督学习。

然而，既然算法中涉及了无监督学习，那么从聚类的数目到聚类中心点的选定，都可能存在误差而导致产生局部最优解，使结果不可信。在数据抽象化的过程中还涉及到虚拟变量，虚拟变量的选取也会对最终结果造成较大影响。

（二）影响结合算法可信度的因素

影响结合算法可信度的因素主要有三个。分别是数据本身、数据预处理以及非监督学习部分的值的选取。

如果数据本身较少而且分布不均，最终结果可能并不能为更大的集体预测规律而服务，得到的结果只和部分人相关。在本论文中，问卷填写者有半数都是本科二年级学生，因此最终的预测结果可能只和本科二年级符合的程度较高，与其他年级的数据模拟会产生偏差。

在数据预处理过程中产生影响主要发生于虚拟变量的选取过程。过多的虚拟变量会使整个模型十分复杂，但现实预测中必然存在较多定类变量。平衡模型复杂度和定类变量重要程度就变得十分重要。

在聚类算法实现初步，人为选取集群和聚集中心点产生错误的概率较大。若全部数据并没有明确聚集程度，改变初始聚集中心点依旧可以得到某些结论，最重要的是可能会使影响最大的特征量发生更替。

（三）结合算法的改进与展望

对于数据本身来讲，最好的办法就是扩大数据规模，尽量使填表人数类别较平均，这样更有助于用样本正确估测整体。在SPSS中可以直接设置虚拟变量并利用logistic回归监督学习。面对线性回归时或大量数据时，应尽量减少虚拟变量的设置，更加注重真实数据的重要性。针对K-means算法容易得到局部最优解的这个问题，目前已有许多相关改进方法的研究，如K-means++、p-swap等。

与传统K-means算法相比，K-means++算法改善了初始化环节，K-means算法的初始聚类中心选取完全随机，因此常由于初始中心选取不具有代表性，影响算法最终结果。与之相比，K-means++算法认为，初始中心选取不应该完全随机，应使初始中心分布尽可能分散，使新的聚类中心与已有中心相距较远，同时为了避免算法直接选取到最远的异常数值点，选择采用概率的方式予以折中。为此，在初始化阶段，距离已有中心较远的点应当有更大的概率被选为聚类中心。

对于k个聚类中心，K-means++算法具有这样的迭代流程：



则集合C即为初始聚类中心集。其中表示数据x到集合C中距它最近点的距离。[2]

在未来大数据时代，数据的获得会更加容易，这将有助于所有的机器学习算法的研究和开发。除直观分析以外，统计学还包括很多高级分析方法。统计学和机器学习、大数据的联合跨学科发展，将成为信息时代的主流。K-means 问题的研究趋势将会朝着更稳定、准确方向发展，随着新的分析技术的发展，K-means 问题历久弥新。

结合算法是对监督学习和非监督学习的综合运用，目前还在实验阶段，并不成熟，应用的算法也是初级且简单的。在机器学习领域，新的理论结果依然在不断出现，与现实问题和数据相关的实用算法也定层出不穷。

致谢

历时一月我们终于完成了论文，在这段奋斗的历程中，我们的知识得到了极大的拓展，实践水平得以提高，小组合作能力更上了一个新的层面。在为完成论文而喜悦时，论文写作的点点滴滴浮现在我们眼前，在这过程中，我们遇到了无数的艰难险阻，架构混乱、数据不足、结果芜杂……都在同学和老师的帮助下成功度过了难关。在此向指导老师表示最诚挚的感谢！

同时，也要感谢我们所引用的论文的各位著者，如果他们的研究成果的启发和帮助，我们将无法完成本篇论文的最终写作。至此，也要感谢朋友和同学们，他们在我们写论文的过程中提供了很多有用的素材和数据，也在我们走进死胡同时帮助我们。

我们的学术水平和所获数据仍有欠缺，所写论文难免有不足之处，恳请各位老师和同学批评指正！

参考文献

[1]张润,王永滨.机器学习及其算法和发展研究[J].中国传媒大学学报(自然科学版),2016,23(02):10-18+24.

[2]任远航.面向大数据的K-means算法综述[J/OL].计算机应用研究, 2020,04(27):1-7.

[3]许心炜,蔡斌,向宏,桑军.基于同态加密的多分类Logistic回归模型[J].密码学报,2020,7(02):179-186.

[4]刘丽艳,朱成全.机器学习在经济学中的应用研究[J].天津师范大学学报(社会科学版),2020(02):51-58.

[5],,Beyond the Artificial Intelligence Hype: What Lies Behind the Algorithms and What We Can Achieve.[J]. Journal of thoracic imaging,2020,35 Suppl 1.

[6] 金圣涵,殷然.国产智能手机境外市场营销策略研究[J].技术与市场, 2019, 026(008):189-190.

[7]刘建花.K-means聚类算法的改进与应用[J].太原师范学院学报(自然科学版),2020,19(01):81-83.

[8] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281–297. MR 0214227.0214.46201. Retrieved 2009-04-07.

[9] Steinhaus, H. (1957). "". Bull. . (in French). 4 (12): 801–804. MR 0090073.0079.16403.