浙江理工大学

**实 验 报 告**

课程名称 数据与知识工程

实验项目名称 基于数据挖掘的淘宝卖家分析

实验学生姓名 孟国庆，温志华，黄梓峰

刘卫春，胡明豪，贺文韬

目录

[一、引言 3](#_Toc106219916)

[二、理论基础与技术要求 3](#_Toc106219917)

[2.1离群点概述 4](#_Toc106219918)

[2.2传统离群点检测方法 4](#_Toc106219919)

[2.3.聚类与K-means 4](#_Toc106219920)

[三、研究方法与手段 5](#_Toc106219921)

[（1）数据集 5](#_Toc106219922)

[（2）数据预处理 6](#_Toc106219923)

[四、结果与讨论 8](#_Toc106219924)

[五、结论与展望 10](#_Toc106219925)

摘要：在当今互联网发展驱动下，淘宝、天猫等电商平台行业发展迅猛，用户规模持续增长。越来越多的人选择方便快捷的网上购物，淘宝作为国内最大的电商交易平台之一，拥有巨大的流量优势。在淘宝APP上，商品的搜索、分类导航、个性化推荐和广告等通过让用户在短时间内了解到更多可能需要或者喜欢的商品而促进购买，同时可以通过数据埋点，收集和分析用户的数据，实现更精细化的运营。本次实验报告基于给定的部分淘宝数据集，使用分析工具jupyter notebook,利用python相关的matplotlib、numpy、pandas库对给定部分淘宝数据集的卖家进行分析，得出可能存在的业务问题，并提出推荐的解决策略。同时，异常卖家行为检测对电商平台卖家公平竞争具有重要意义。本次实验报告基于某企业的卖家销售数据、店铺评分数据、买家评分数据，通过数据预处理方式得到处理好的数据表，在此基础上通过统计分析方法，k-means算法对异常卖家进行聚类分析，从而得到若干有益的结论。

# 一、引言

目前大量的电子商务平台充斥着人们的生活当中，并且有关 的物流、销售等等产业在也在不断的完善。现在对于年轻人来讲， 网络购物已经成为最为常见的购物形式，并且它在经济社会发展 当中也具有相当大的作用，已经成为一股不可忽视的力量。对于 电商用户来讲，只需要点击鼠标就可以完成对于商品浏览，以及 之后的购买和售后。因此购物不再会受到时间和空间的限制，更 多的人乐意接受这种网络购物的行为。也正因为如此，消费者的 购物理念和行为发生了翻天覆地的变化。因此对于电商来讲，需 要进一步的研究用户的行为，通过数据挖掘的方式来对商品的销 售进行预测，并且对自己的商业模式与相关的商业系统进行调整， 以迎合用户的选择，从而为企业创造更加巨大的经济利益。

# 二、理论基础与技术要求

离群点被定义为一个显著不同于其他数据分布的数据对象，通过分析离群点数据分布特征，可以从海量数据中挖掘异常信息、提取兴趣模式等。因此离群点检测(outlier detection)成为数据挖掘领域的研究热点之一。离群点检测目的是通过数据挖掘方法找出不同于大规模数据中的异常点，并发现潜在的、有意义的知识。

传统离群点的检测方法众多，一般经典的离群点检测方法通常分为基于统计学的、基于聚类的、基于分类的和基于邻近性的方法四大类。

## 2.1离群点概述

离群点可以定义为如孤立点、异常点、新颖点、偏离点、例外点、噪声、异常物等，即一个显著不同于其他数据分布的数据对象。产生离群点的原因以及挖掘离群点大体分为:

a) 数据来源异常，这类离群点往往意味着异常情况的产生，例如不寻常的信用卡消费记录，往往意味着诈骗行为的发生。对于这类离群点，其产生的原因明确，挖掘此类离群点可以帮助人们及时处理异常情况。

b)其外在表现同样为异常数据的产生，所不同的是其产生的原因具有很高的学术研究价值，属于主观上受欢迎的离群点。对于这样的离群点，其发现的数量越多越具有学术价值。例如绝症的罕见自愈等，通过对其原因的研究可以帮助人们加深对该疾病的认识。

c)数据测量和收集误差，主要是由于人为错误、测量设备故障或存在噪声。由于这类离群不提供有趣或有用的信息，只会降低数据和其后数据分析的质量，所以目标是消除这类离群，如传感器故障造成的误报警等。

## 2.2传统离群点检测方法

**（1）基于统计学的离群点检测方法**

离群点检测最早出现在统计领域，基于统计学的离群点检测方法的一般思想是: 以给定数据集的生成模型为前提条件，然后将该模型中低概率区域中的对象判定为离群点。如何得到数据集的生成模型，一般分为参数法和非参数法。

**（2）基于分类的离群点检测方法**

对于具有类标签的数据集，可以训练一个区分正常数据和离群点的分类器。然而对离群点检测问题，数据集往往高度有偏。虽然可以使用现有的分类算法检测离群点，但离群点训练样本数严重不足，很难构建一个准确率较高的分类模型。因此，通常的做法是构建单分类模型(one-class model) ，即只使用正常数据训练模型，这样不属于正常类的数据点判定为离群点。在一些具体应用中，也可能将正常数据分为几个类，把不属于这几类的数据判定为离群点。传统分类模型如 C-SVM、SVDD和KNN。

**（3）基于分类的离群点检测方法**

基于邻近性方法的核心思想是定义出数据之间的邻近性度量，并根据此度量的值判定离群点。其中比较典型的方法是基于距离的方法以及基于密度的方法，前者以距离体现邻近性，后者以密度体现邻近性。

**（4）基于聚类的离群点检测方法**

离群点检测一般通过无监督学习获取数据分布特征，因此常见的聚类算法在修改后都能应用于离群点检测，其方法主要通过考虑对象与簇之间的关系检测离群点。比较经典的算法有K-means、DBSCAN、CLAＲANS、CHAMELEON、BIRCH、STING、WaveCluster、CLIQUE和FindCBOLF等。

## 2.3.聚类与K-means

与分类、序列标注等任务不同，聚类是在事先并不知道任何样本标签的情况下，通过数据之间的内在关系把样本划分为若干类别，使得同类别样本之间的相似度高，不同类别之间的样本相似度低（即增大类内聚，减少类间距）。

聚类属于非监督学习，K均值聚类是最基础常用的聚类算法。它的基本思想是，通过迭代寻找K个簇（Cluster）的一种划分方案，使得聚类结果对应的损失函数最小。其中，损失函数可以定义为各个样本距离所属簇中心点的误差平方和：

其中代表第个样本，是所属的簇，代表簇对应的中心点, 是样本总数。

**算法具体步骤**

K-Means的核心目标是将给定的数据集划分成K个簇（K是超参），并给出每个样本数据对应的中心点。具体步骤非常简单，可以分为4步：

数据预处理。主要是标准化、异常点过滤。

随机选取K个中心，记为

定义损失函数：

令t=0,1,2,... 为迭代步数，重复如下过程直到收敛：

对于每一个样本，将其分配到距离最近的中心

对于每一个类中心k，重新计算该类的中心

**K-Means最核心的部分就是先固定中心点，调整每个样本所属的类别来减少；再固定每个样本的类别，调整中心点继续减小。两个过程交替循环，单调递减直到最（极）小值，中心点和样本划分的类别同时收敛。**

# 三、研究方法与手段

## （1）数据集

我们的实验用到的数据集来源于淘宝。该数据集中包含了淘宝2010年三月份到七月份的，大概几百万的数据。

该数据集主要包含了user表、描述卖家店铺星钻冠等级、开店时长、总交易笔数相关信息；rate表描述了卖家店铺评分信息；trans表描述了卖家店铺开店时长、总交易笔数、价格、数量总的金额、买家ID相关信息；dsr表描述了每个订单中买家对于买家的服务得分、发货得分、物流得分、商品得分。

实验数据的4个表格，user表、trans表、rate表、dsr表，以下是每个表格的字段：

user表：

* 卖家ID
* 卖家评级
* 开店时长
* 销售量

trans表：

* 卖家ID
* 开店时长
* 总交易笔数
* 价格
* 数量
* 总支付费用
* 买家ID

rate表：

* + 卖家ID
  + 评分：评分标准为-1/0/1，分别代表差/中/好评
  + 日期：日期格式为日-月-年 时:分:秒

dsr表：

* + 卖家ID
  + 买家ID
  + 日期：日期格式为日-月-年
  + 服务得分：离散属性。数值为1到5分
  + 发货得分：离散属性。数值为1到5分
  + 物流得分：离散属性。数值为1到5分
  + 商品得分：离散属性。数值为1到5分

## （2）数据预处理

数据的预处理是指对所收集到的数据分析之前做的排序、筛选、排序、删除异常值等必要的处理。数据预处理一方面是为了提高数据的质量，另一方面也是为了适应所做数据分析的软件或者方法。数据预处理的步骤主要有数据清洗、数据集成、数据变换、数据归约四大步骤。

下面是介绍每个表的处理方式：

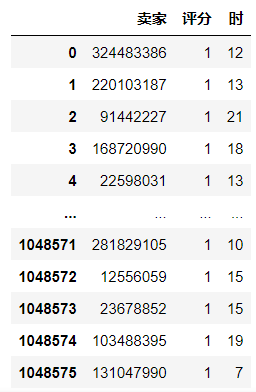
dsr表日期中提取出月份，对表格中缺失值用总平均数进行填充。如图1-1所示

表格

描述已自动生成

图表 1-1

rate表中提取出时，用小时作为时间单位。如图1-2所示：



图表 1‑2

trans表中计算出差额（差额=|总支付费用-价格\*数量|），然后根据卖家ID进行分组，计算出每个卖家的差额总值作为异常值，并作为新的特征，合并到user表中，得到新的user表。对新的user表处理中，对店铺等级用数字进行替换，对异常值中的缺失值用0进行补充。如图1-3所示：

表格

描述已自动生成

图表 1‑3

通过分析dsr表中的物流得分字段，我们可以通过每个月份的评分高低来判断该月份的物流情况，对于评分较低的几个月份我们可以适当增加物流投入。之后我们可以结合trans表中的总金额，来分析对于物流评分较低的月份投入更多的物流成本是否值得，从而最终决定要不要增加物流。

通过分析分析一天中每个时间段的评分次数高低来决定在高概率评分的事件增加客服或者进行产品的推广服务。

通过离散点检测分析异常卖家数据，进而判断卖家的行为是否正常。

# 四、结果与讨论

1.实验得分分析的三个结果，即物流得分分析、评分表分析和异常卖家行为分析如下：

（1）物流得分分析

将每个月的物流得分进行汇总和绘制，如图2-1所示：

图表, 折线图

描述已自动生成

图表 2‑1

从月份角度进行分析，8月的平均得分最低，12月的平均得分最高。因此可以在8月-10月，电商卖家可以对物流的发货速度进一步提高，弥补物流运输中速度慢的问题。

（2）评分表分析

对一天中每个小时段中评分次数的汇总，可以得到评分次数与时间的关系，如图2-2所示：

图表, 折线图

描述已自动生成

图表 2-2

从时间角度分析，在上午十点至晚上十点，进行评分的人数居多，在晚上七点至九点达到顶峰，因此可以在该时间段对用户进行相关产品的推送、增加客服数量等措施。

（3）异常卖家行为分析

异常卖家数据的分析可以使用离群点检测的方法，在离群点检测的方法中主要包括基于统计、基于邻近度、基于密度和基于聚类。该实验是通过使用聚类中K-means算法进行聚类分析，聚成两类，一类是正常的卖家行为，一类是异常的卖家行为，聚类之后再对每个类的特征进行分析，每个类的特征分布如图3-1所示：

图表, 直方图

描述已自动生成

图表 3-1

图3-1中，类别=1的为异常类，对异常类中的特征进行分析可以得出结论：销售评级大于10，销售时长低于1500小时，总交易笔数小于20万，异常值在30万~70万可以判定卖家的数据存在异常。

**2.讨论如下：**

（1）物流得分分析可以得出：一年中物流需求不是很稳定，从月的角度看上半年较稳定，但由于下半年备战双十一和过年，物流需求大大增加。要明确：物流路线的成本控制，是整个物流体系中最重要的一环，但物流路线的设计，和货品流量，卡车费用，运输费用这些因素都息息相关。也会由于某些产品的特殊性，而选择以更昂贵的价格来达到更快的运输速度。

所以从8到10月商家应准备和指定对策，如商品整合（aggregate)以及多点停留((multi-stop)策略、灵活利用统筹点 （Pool points)以及分批使用越库功能 (Cross-docking)等。

（2）对于评论流量分析，在上午十点至晚上十点，进行评分的人数居多，在晚上七点至九点达到顶峰。对于晚上七点至九点是高峰，就要做出对策，来防止网站在流量高峰期崩溃。例如：

* CDN加速：除了提高起重设备的性能，更有效和常用的方法是CDN加速度。CDN加速的原则是将网站的内容分布到整个网络的加速节点。用户可以通过不同的节点获取附近所需的内容，提高终端用户访问网站的响应速度。同时可以减轻服务器的压力，分散对加速节点的访问量，从而提供服务器的稳定性。
* 设备性能：服务器带宽和网络设备是决定因素。充分的网络带宽和设备性能保证了网站的稳定性。但是高性能通常意味着高成本。为了避免不必要的浪费，可以通过之前的数据预测流量，选择满足需求的配置。
* 静态页面：静态页面也会在一定程度上对环境服务器造成压力。动态页面需要读取和调用数据库内容来显示数据，增加数据库读取的数量，占用服务器资源，影响网站的速度。利用网站进行静态化，直接删除浏览数据库的操作，减少链接，网站可以处理的访问量也可以提高。

异常卖家行为分析的对数据异常进行界定：日常工作中的流量数据、用户数据、转化数据、交易数据都有固定的波动周期，每个周期内的数据变化应该是趋于稳定的，如果某天的数据不在符合预期的范围内稳定变化就是数据异常。

实验中的异常检测方法：我们使用历史数据来构建由每个被监测的数据的估计值。将实时数据与这些值进行比较，并分配一个分数。基于从最近的数据观察得到的阈值，决定实时数据是否为异常。这种方法的优点是阈值不是静态的，而是实时的。

通过实时的异常数据监测，我们可以清楚地看到网站流量的差异，在发生异常状况时迅速进行故障排除和修复，减少网站停机，减少潜在客户的流失。

# 五、结论与展望

在日常数据分析工作中，数据异常是很常见的现象，我们可以通过对自身业务的理解来觉察异常数据，然后再界定与定位异常问题，其次通过业务拆解法、指标推进法和用户行为路径法来提出假设，最后运用对比与细分的方法来验证。遵照这个流程，一般能够找到数据异常的深层原因。

当发现数据异常时或者接到数据异常分析任务时,我们可以联系产品相关的信息,在范围维度(自身、竞对、大环境)和内容维度(产品、技术、用户、运营)结合给出合理的猜测,然后通过查看一些大环境变化数据或者细分的产品数据来验证我们的猜测。遵照这个流程,一般能够找到数据异常的深层原因,当然,着需要花费大量的时间和足够的耐心,但它能够让我们更深更全面的了解自己负责的产品的相关信息,并为未来的产品决策提供指导。