基于地缘特征的相似人群广告投放算法

算法流程说明书

目录

[一、 概述 2](#_Toc512637632)

[二、 数据标引流程 2](#_Toc512637633)

[1、 购买力特征采集 2](#_Toc512637636)

[1.1 算法概述 2](#_Toc512637638)

[1.2 数据采用 3](#_Toc512637639)

[1.3 数据量化 3](#_Toc512637640)

[1.4 总体结构 4](#_Toc512637641)

[1.5 结果预期 5](#_Toc512637642)

[2、 年龄层特征采集 6](#_Toc512637643)

[三、 广告系统冷启动与相似性投放 6](#_Toc512637644)

[1、 冷启动特征 7](#_Toc512637647)

[1.1 广告特征 7](#_Toc512637648)

[1.2 群体特征 7](#_Toc512637649)

[2、 相似性投放 7](#_Toc512637650)

[2.1 欧几里得距离 7](#_Toc512637652)

# 概述

基于地缘特征的广告（即社交广告）是当前大环境中最普遍的广告种类。该种类广告依托人群本身的相似性进行广告投放，也是最常见的广告投放形式，如海报、传单广告框等均是依据地缘特征进行投放的一种方案。

这种方案不同于基于互联网的广告方案可以有效的采集到群体中的个人特征，基于群体整体特征的投放方法要求高精度的群体特征采集，从而达到最大的投放成功率。

在以社交群体为基础的人类社会中，群体特征的最本质特征便是地缘特征，住在高档小区的人大概率会比住在普通名居房中的人群拥有更强的购买力或者追求更高的生活品质，这也是我们投放算法的启动关键。

在获得人群的地缘特征后，我们将会对广告投放系统建立冷启动标准，在后续执行过程中不断改进，取得更加精确的算法结果。

本算法将分为两部分进行，一部分是关于地缘特征的采集算法，另一部分是广告的相似性投放算法。

# 数据标引流程

利用城市大数据多元融合来对地段人群进行区分在当代极具价值，它可以有效衡量一个地段的整体特征。在纽约时代广场的女性必然大概率会比来自偏远山区的女性更能消费得起高档化妆品，这种地缘特征也为广告投放商提供了参考，这也是在不同地段见到不同广告的根本原因。

然而传统的广告投放方式具有太多不确定性，如地段内的年龄分布、阶层分布均是无法直接获取的数据，而本算法目标便是通过当前大数据处理方案，通过机器学习的方式挖掘出导致不确定性的本质特征，提供稳定的地段特征信息，对地段内人群的基本购买力进行划分，从而实现广告的精准投放。



## 购买力特征采集



### 算法概述

地段价值的高低也是评估购买力的一个有效标准。面对大笔金额差异，处于上层精英社会的人群相比普通老百姓大概率住在更加优质的地段。

地段价值进行评估在当前时代极具价值，比如买房，门面选址等等，决定地段价值的因素也越来越显著可寻。对于选址来说，在经济允许的条件下，更专注关于地段相关的特征选取，通过对地段周边各类设施配套水平，交通，商场等进行学习。

本算法则更偏向于高效投资，这也是对大多数人而言更关注的地方，地段的实际价值不容易准确评估量化，但地段与地段之间的好坏比较是显著的，在不考虑政策相关因素时，通过已经得到的数据将地段划分为5~10类，从低价逐步到高价的类型。而后通过深度学习训练得到每类地段周边数据模式，进而找出最具购买力的地段。

算法的总体逻辑是将地段根据房价划分为5~10类，将地段周边的数据量化分层输入深度学习模型进行训练，进而找出影响地段价值变化的关键因素。然后将所有已知城市数据量化输入，进而找出最具投资价值的地段。

### 数据采用

首先为了对地段价值进行评估，需要了解其周边各类设施配套水平，如交通，商场，学校等等，这些数据可以通过路网，POI结果中得到。

其次数据获取后需要被二次区分，优质商场或者学校对于地段的影响远大于普通商场与学校。

然后数据需要被合理量化，更标准一致的数据集可以更好的影响最终训练结果。

所需必须数据如下：

1.连续时间短的的房价信息（用于数据标引）

2.路网数据（道路，车站，地铁）

3.河流，山脉等地理数据

4.POI数据（商场，学校等）

扩选其他数据包括

1.人流数据（滴滴，公交）

2. PM2.5数据（间接衡量交通、工业发展程度）

…

### 数据量化

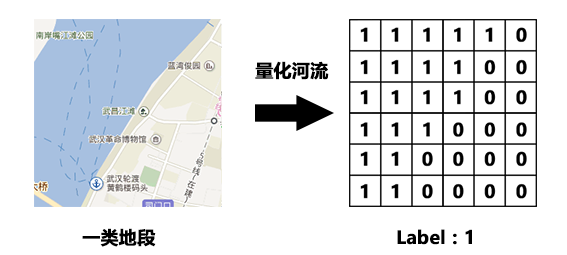
1. 以被标引的数据为中心，将周边2公里范围离散化

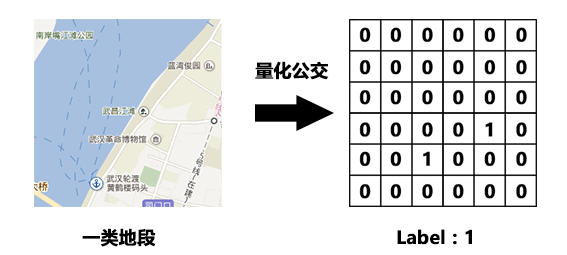
离散化数据可能会损失部分有价值的细粒度数据，但同时带来的好处是可以直接准确的评估数据的影响程度。比如商场POI对结果的影响，甚至商场POI的具体面积，占地结构对结果的影响。

1. 为数据项标引

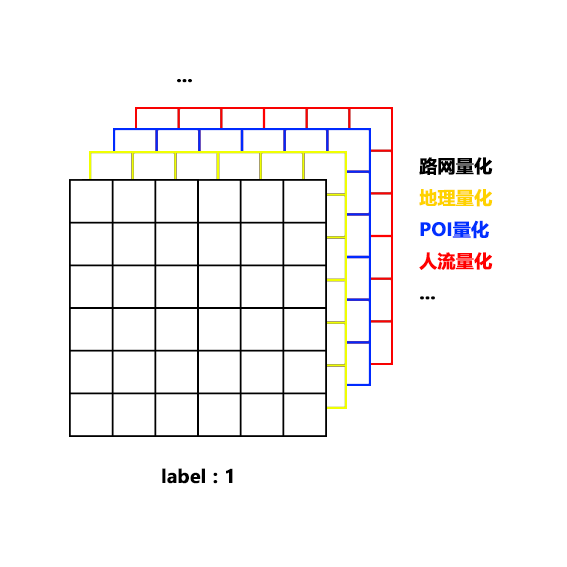
启动前将对各种不同的地段数据集进行标引，确定数据集属于某种具体的分类，通过已有的房价信息将地段划分为1、2、3…等多个等级

1. 将多元数据分别映射至各个离散区域中，形成量化数据

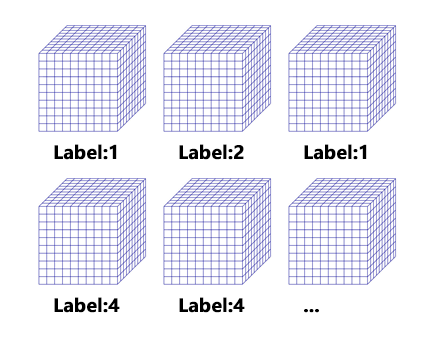




1. 得到标记样本的数据集



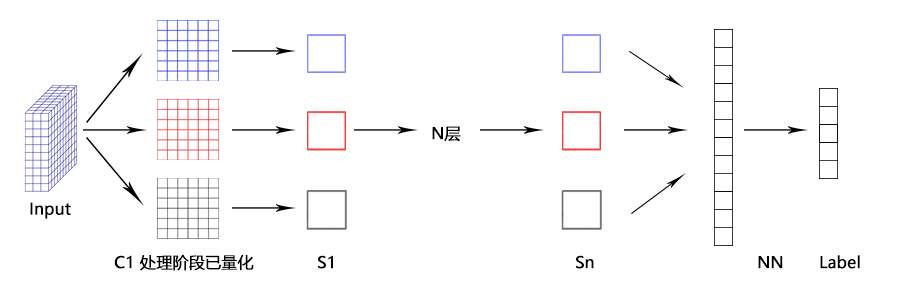
1. 得到拥有一组样本的训练集



### 总体结构

总体结构为：

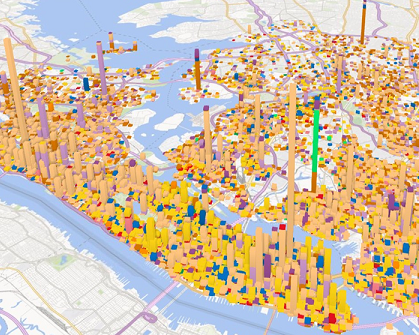
1. 训练过程：将80%已标记地段样本的量化数据分离到C1状态（分离出路网，地理信息，POI，人流等量化特征），由多层感知器提取其最显著特征到Sn，在NN阶段将特征向量重新连接，计算调整特征输出权重直到标记样本用完或趋近收敛。
2. 验证过程：将20%已标记地段样本的量化数据作为输入，通过NN计算得到Label，计算准确率与召回率。
3. 应用过程：将城市所有未标记数据作为输入，计算得到城市地段价值Label。



1. 输入层：输入为带标签样本量化后的数据集（分层表示路网，地理信息，POI，人流等）
2. C1：将量化数据集拆分出有效特征（C1在处理阶段已量化，将每层拆分）
3. S1：二次滤波提取有效特征
4. Sn：n次卷积后的结果数据集
5. NN：映射为线性向量后的神经网络分类
6. Label：最终决定分类的Label

### 结果预期

计算得到全城的地段价值分类，可以显著的发现最具购买力的地段。



## 年龄层特征采集

算法的总体逻辑是计算出周边老年会所、养老院、幼儿园、中小学、娱乐会所等典型标志性年龄划分的POI为基础通过POI的数量，计算取得模糊的年龄层分布

# 广告系统冷启动与相似性投放



相似人群拓展的目的是基于广告主提供的目标人群，从海量的人群中找出和目标人群相似的其他人群。在实际广告业务应用场景中，相似人群拓展能基于广告主已有的消费者，找出和已有消费者相似的潜在消费者，以此有效帮助广告主挖掘新客、拓展业务。

相似人群拓展基于广告主提供的一个基础特征人群，自动计算出与之相似的人群（称为扩展人群）。

## 冷启动特征

### 广告特征

广告的基本信息（广告名称、投放时间等）

广告的推广目标

广告的标签

投放平台

投放的广告规格

所投放的广告创意

广告的受众

广告出价信息

目标购买力

目标年龄层

目标POI

### 群体特征

地段购买力

地段模糊年龄层

地段周边POI

## 相似性投放

推荐系统需要数据作为支撑。但在刚刚开始做推荐的时候，是没有大量且有效的用户行为数据。这时候就会面临着“冷启动”的问题。没有用户行为数据，就利用商品本身的内容数据。这将是算法前期的做法。



### 欧几里得距离

根据广告与群体之间的共同特征为维度，建立一个多维的空间，对单一维度上的评价组成的坐标系即可定位群体在这个多维度空间中的位置，那么任意两个位置之间的欧式距离在一定程度上反应了广告与群体之间的相似程度。

就其意义而言，欧氏距离越小，广告与投放群体相似度就越大，欧氏距离越大，广告与投放群体相似度就越小。

1. 冷启动数据系统
   1. 数据采集：web爬虫爬取搜房网、房天下等网站的房地产价格数据
   2. 数据清洗：对过时数据，失效数据通过正则匹配等方式进行清洗
   3. 数据分析：
      * 目的：通过与周边相关性评估缺失数据的小区房产价格
      * 特征提取：房价、商圈密集度、地域
      * 训练集：已知房价以10%挖空作为训练集
      * 测试集：挖空数据做测试集
      * 算法选择：KNN，贝叶斯等（择优）
2. 广告位推荐系统
   1. 定价模型
      * 根据已知小区房价水平，似然评估其消费水平，不要求绝对准确，用作推荐系统冷启动数据基础
      * 根据广告位所在小区房价水平和周边商圈密集程度，为广告位价格做初始评估
   2. 推荐组合模型（确定投放方案）
      * 评价标准：广告商资本量，投放产品领域，投放产品消费档次，及广告位价格
      * 目的：确定广告商在各个消费群体的投放比例和地域分布
   3. 协同过滤模型
      * 用户特征采集：注册过程
      * 用户行为采集：操作过程（浏览，购买等）
      * 相似性分析：向人物画像系统提供数据取得相似性信息，为推荐组合模型提供精准投放方案
3. 人物画像系统
   1. 特征提取
      * 特征筛选：主要考虑投放商领域，偏好价位，偏好地域等
   2. 相互推荐
      * 计算用户相似性
      * 计算广告位相似性
      * 计算推荐
        1. 为相似投放商提供相似方案
        2. 将相似广告位推荐给购买过该类广告位的投放商