|  |  |
| --- | --- |
| 密级： | 公开 |
| 版本： | V1.0 |



潜在需求指标设计文档

2016年2月2日

**版权声明**

本使用手册版权由百分点信息科技有限公司所有，未经百分点信息科技有限公司的许可和授权，任何机构及人员不得将本手册及本使用手册包含的任何文字、内容及图片以任何理由，任何方式或手段（电子的或机械的）进行复制或传播。

© 百分点信息科技有限公司版权所有 保留所有权利 2015年9月

本手册可能涉及百分点信息科技有限公司的专利（或正在申请的专利）、注册商标（或正在注册的商标）、版权或其它知识产权，除非得到百分点信息科技有限公司的明确书面许可协议，本手册不授予使用这些专利（或正在申请的专利）、注册商标（或正在注册的商标）、版权或其它知识产权的任何许可协议。

北京百分点信息科技有限公司

网址：www.baifendian.com

地址：北京市海淀区北四环西路66号，中国技术交易大厦A座15层

电话：+86-10-82484868

传真：+86-10-82484869

**修订记录**

|  |  |  |  |
| --- | --- | --- | --- |
| 版 本 | 修订说明 | 编制/日期 | 审核/日期 |
| V1.0 | 潜在需求指标设计文档 | 付宇/  2015.11.25 |  |
|  |  |  |  |
|  |  |  |  |

目 录

[一、 概要说明 1](#_Toc442200766)

[1.1 目的 1](#_Toc442200767)

[二、 整体流程设计 1](#_Toc442200768)

[2.1 整体功能模块图 1](#_Toc442200769)

[2.2 功能模块详情介绍 2](#_Toc442200770)

[2.2.1 黑洞系统 2](#_Toc442200771)

[2.2.1.1 用户行为数据提取 2](#_Toc442200772)

[2.2.1.2 用户画像数据提取 2](#_Toc442200773)

[2.2.1.3 拼接 3](#_Toc442200774)

[2.2.2 数据预处理模块 3](#_Toc442200775)

[2.2.2.1 数据过滤模块 4](#_Toc442200776)

[2.2.2.2 数据Label标定模块 4](#_Toc442200777)

[2.2.2.3 数据离散化处理模块 4](#_Toc442200778)

[2.2.2.4 数据划分模块 5](#_Toc442200779)

[2.2.3 规则分析模块 5](#_Toc442200780)

[2.2.3.1 规则挖掘模块 5](#_Toc442200781)

[2.2.3.2 人工分析模块 5](#_Toc442200782)

[2.2.3.3 评价模块 6](#_Toc442200783)

[2.2.4 特征工程 6](#_Toc442200784)

[2.2.4.1 单特征分析模块 6](#_Toc442200785)

[2.2.4.2 特征提取 6](#_Toc442200786)

[2.2.4.3 特征选择 7](#_Toc442200787)

[2.2.5 模型构建 7](#_Toc442200788)

[2.2.5.1 模型训练模块 7](#_Toc442200789)

[2.2.5.2 模型选择模块 7](#_Toc442200790)

[2.2.6 结果预测模块 8](#_Toc442200791)

[2.2.6.1 规则模块 8](#_Toc442200792)

[2.2.6.2 模型 8](#_Toc442200793)

[2.3 数据流图 9](#_Toc442200794)

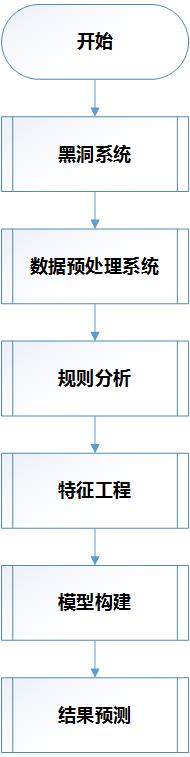
# 概要说明

## 目的

根据已有的用户购买行为以及历史累计的用户画像数据，预测用户在未来三个月（一季度），在电商品类下的潜在需求

# 整体流程设计

## 整体功能模块图



## 功能模块详情介绍

### 黑洞系统

#### 用户行为数据提取

黑洞系统需要提供指定日期的（按月份）用户在指定品类下的用户行为记录，提供的数据格式为:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| gid | type | action | level | category | l\_date |
| 用户唯一标示 | business/media,  品类类型 | pageview，  addcart,  order,  search等，用户行为类型 | 品类对应的类目层级 | 品类名称 | 时间戳（截止到月份,201505) |

将其对应的结果保存在冷集群的HDFS上。

#### 用户画像数据提取

打平后的用户画像数据，提供的主要维度包括用户的人口属性、上网特征、营销特征、长期购物偏好以及长期兴趣偏好，将其保存冷集群的HDFS上。

#### 拼接

以gid字段作为关联， 按照配置文件指定获取字段，利用MR程序输出最终的数据集结果。配置文件可以指定提取的字段类型（business/media/profile）、以及行为类型。最终输出的结果是一个数据矩阵，其具体表示如下所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| gid | column1 | column2 | column3 | … … | columnN |
| 用户的唯一标示 | 电商/媒体/用户画像的维度，具体值可以其对应的权重 | 电商/媒体/用户画像的维度，具体值可以其对应的权重 | 电商/媒体/用户画像的维度，具体值可以其对应的权重 |  | 电商/媒体/用户画像的维度，具体值可以其对应的权重 |

无值的情况下可以设置为0, 其最终输出是一个非常稀疏的矩阵。

### 数据预处理模块

数据预处理模块主要的功能获取优质的数据集。主要包括以下4个方面的内容：数据过滤模块、数据Label标定模块、数据离散化处理模块以及数据划分模块。

#### 数据过滤模块

数据过滤模块主要可以通过配置一些规则删除数据集中存在的脏数据（目前没有一个很好的实施方案，可以暂时），根据需求，手工编写程序进行过滤。

#### 数据Label标定模块

该模块主要的目的是为已有的原始数据标定相关的类标号，目前的方式可以分为两种：人工标注以及规则标定两种。

规则标定是指用来设定规则，使用程序自动为数据集标注相关的类标号，适合于规则明确，数据量较大的情况使用。

人工标注是需要人手动对于数据集的类标号进行标注，适合用于规则不明确，数据量较少的情况下使用。

采用的规则是尽量使用规则标定，如果数据中的正例过少，可以使用人工标注的方法增加正例个数。

#### 数据离散化处理模块

该模块主要是对原始数据进行去纲，进行归一化处理，保证各个维度的区间保持一致。同时针对不同的应用场景，可以采用One-Hot Coding或者Conditional Probability Coding等策略对于特征的取值进行处理。

#### 数据划分模块

该模块的目的是为了保证两个目的：一个是训练集中的正负样例比例可以保持一定的比例，而不应该出现偏移（负例过多，正例很少）；同时另一个目的是保证测试集中的正负样例比例与线上的实际环境尽量保持一致或者近似。因此该模块针对训练集和测试集提供一个可配置文件，可以设置数据集的大小以及正负比例。

同时应该提取抽样程序，针对数据量非常大的情况可以使用抽样程序，获取指定数据量的抽样数据集。

### 规则分析模块

规则分析模块主要是采用的策略是“模式挖掘+人工挑选”的方式确定相关的规则集合，在对规则进行筛选的时候，应该采用相对比较严格的筛选条件，已达到规则集中的规则判断的结果可信性较强。主要包含的模块包括：人工分析模块、规则挖掘模块、评价模块以及输出模块。

#### 规则挖掘模块

利用Spark MLlib以及其他的规则挖掘模块（主要采用关联规则模块）获取对应的关联规则。

#### 人工分析模块

可以通过人工分析的手段获取相关的具有区分价值的规则，将其作为最终结果输出。

#### 评价模块

可以采用张琦的Venice模块，设定相关的评价指标（F因子，精确率，召回率）等，评估相关的规则，可以将对应的阈值设置相对较高范围（相对于模型），从而保证其输出的结果能够具有较高的可信度。

### 特征工程

特征工程主要是针对原始特征(low-level)特征进行分析，通过特征提取以及特征选择获取对应的最优特征集合。使用特征工程可以使其训练模型减少训练时间，提高模型的泛化性。

#### 单特征分析模块

该模块主要针对单个特征与类别属性关系，以及相关的分布情况进行分析，主要分析的指标包括两个：单个特征与类别属性的相关性以及单个特征的覆盖情况。可以人工进行分析，也可以为后来的特征提取与特征选择提供辅助信息。

#### 特征提取

该模块主要通过单特征的分析结果，通过不同的构造方式获取新特征（High-level)，可以通过lasso以及卡方的方法获取相关指定的特征集合。

#### 特征选择

使用PCA的方法对于已有的特征集合进行降维处理，从而保证获取最终模型对应的特征集合。

### 模型构建

当数据量较大的时候可以采用Spark平台进行模型训练，在训练过程中可以采用多个模型并行训练的方式，选择在测试集中最优的模型作为上线模型，其中选择的模型应该是工业界中已广泛使用的模型算法，目前暂定LR，CNN，RF，GBDT，SVM等模型进行对比。

#### 模型训练模块

针对数据集的大小以及时间限制，可以采用Python sklearn以及Spark MLlib两种方式训练模型。

#### 模型选择模块

根据各个模型在多个测试集（大于等于5个）上的平均值，选择最终的模型，作为最终的输出模型。其中上面的平均值是指我们测试模型时指定的性能测试指标的在各个测试集的平均值，同时也要满足人为设定的上线阈值。

如果最终输出的最优模型仍然无法满足上线阈值，那么可以从两个方面进行优化：

1) 从特征选择中，添加新的特征（包括Low-level以及High-level)，提高模型的拟合性。

2）优化数据集，优化训练集中的数据，添加正负比例

3）使用集成算法，组合输出模型

### 结果预测模块

在预测模块中，我们采用的是一种类似决策树的规则，使用规则匹配为先，模型预测其次的策略，进行结果的预测。

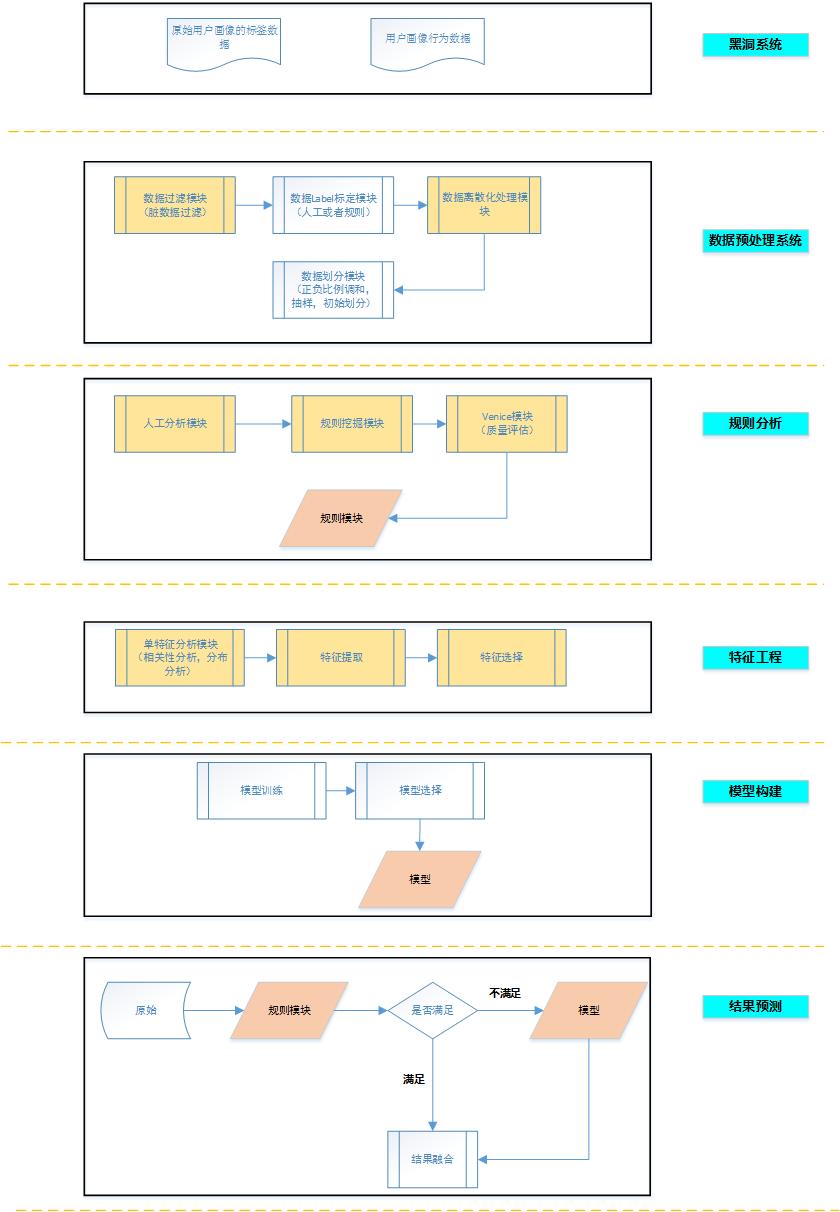
#### 规则模块

规则模块采用的精确匹配的规则，当预测数据符合其中一条规则的时候，就将其结果进行输出，同时计算对对应的权重值，理论上规则模块输出结果的权重值要比模型输出结果的权重值大；如果规则匹配，不进行相关的模型判断。

#### 模型

当规则模块无法进行判断的时候，利用模型进行预测，如果满足阈值条件，则输出对应的结果值，通知计算对应的权重值，理论上模型输出的结果的权重值要比规则输出结果的权重值大。

## 数据流图



上图展示的是潜在需求整体的数据流图，整个过程如下所示：

1）黑洞系统，提供指定的用户行为数据以及用户画像指标数据，同时利用MR程序进行统计融合，最终输出模型需要的原始数据集。

2） 数据预处理模块，该模块的主要功能是优化训练和测试集合，通过脏数据的过滤、特征的离散化、数据类标号的标定以及训练集、测试集中征服样例的比例，达到优化数据集的目的。

3） 规则分析模块，该模块的主要功能是利用“模式挖掘+人工分析”的策略，获取高质量的关联规则模式集合，从而达到利用该规则集合进行预测的目的，该规则的产生标准要较高与模型类的预测标准。

4）特征工程模块，利用特征选择和特征提取获取最优的子特征集合

5）模型构建模块，通过已有的模型库（LR，CNN，RF，GBDT，SVM），训练模型，对于小数据可以采用python sklearn，对于大数据量可以采用Spark MLlib进行训练，同时通过对多份测集（大于等于5份）的测试，获取各个模型在测试集上的平均测试结果，选取最优的模型作为上线模型（同时该模型要满足人为设定的上线阈值）。

6）结果预测模块，采用规则优先，模型其次的策略，对于输入数据进行预测。规则预测采用的精确匹配的策略，满足其中的一条规则，则输出对应的结果，其输出的权重值较大。规则无法预测的数据，采用模型进行预测，当其预测的结果满足人为设定的阈值的时候，则输出对应的结果，其输出的权重较小。