

课程设计报告

**广告点击率预估挑战赛**

**姓 名：**解博棋

**学 号：**19851008

**专 业：**计算机科学与技术

**系 别：**计算机与信息技术学院

**指导教师：**王志海

**2022年6月**

# 广告点击率预估挑战赛课程设计报告

关于广告点击率预估挑战赛，分为训练集和测试集，包括13个特征字段和六天的数据，其中训练集有39万条，测试集有7万条，同时会对部分字段信息进行脱敏，根据这些数据进行数据分析。

## 1 数据分析网站Weka简介

Weka的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis）是用于数据挖掘任务的机器学习算法的集合。它包含用于数据准备、分类、回归、聚类、关联规则挖掘和可视化的工具。同时Weka也是新西兰的一种鸟名，而Weka的主要开发者来自新西兰。

Java机器学习软件，Weka是一组用于数据挖掘任务的机器学习算法。它包含工具，用于数据，准备、分类、回归、聚类、关联规则挖掘和可视化。

维卡鸟只在新西兰的岛屿上发现，它是一种不会飞的鸟，具有好奇的天性。它的发音是这样的，鸟的声音也是这样。Weka是在GNU通用公共许可证下发布的开源软件。

我们已经建立了几个免费的在线课程，使用Weka教授机器学习和数据挖掘。这些课程的视频可以在YouTube上获得。

Weka支持深度学习！

## 2 Weka软件功能介绍

除了利用Weka进行数据挖掘的同时，Weka也很适合用来开发新的机器学习算法。[Weka是根据GNU通用公共许可证](https://www.gnu.org/licenses/gpl.html" \t "https://www.cs.waikato.ac.nz/ml/weka/_blank)发布的开源软件。Weka还支持[深度学习](https://deeplearning.cms.waikato.ac.nz/)。

如果想自己实现数据挖掘算法的话，可以参考Weka的接口文档。在Weka中集成自己的算法甚至借鉴它的方法自己实现可视化工具并不是件很困难的事情。

2005年8月，在第11届ACM SIGKDD国际会议上，怀卡托大学的Weka小组荣获了数据挖掘和知识探索领域的最高服务奖，Weka系统得到了广泛的认可，被誉为数据挖掘和机器学习历史上的里程碑，是现今最完备的数据挖掘工具之一（已有11年的发展历史）。Weka的每月下载次数已超过万次。

2014年3月起，[新西兰怀卡托大学](https://baike.baidu.com/item/%E6%96%B0%E8%A5%BF%E5%85%B0%E6%80%80%E5%8D%A1%E6%89%98%E5%A4%A7%E5%AD%A6/417707" \t "https://baike.baidu.com/item/weka/_blank)将推出Weka免费网课，课程分为初级和高级两个部分，每个部分时长5周。初级课程将于2014年3月3日开课，高级课程于2014年4月下旬开课。课程具体内容参见怀卡托大学网站Weka MOOC。课程在优酷网站也有专辑。

Weka系统汇集了最前沿的机器学习算法和数据预处理工具，以便用户能够快速灵活地将已有的处理方法应用于新的数据集。

2.1 Weka的历史

Weka的历史，1992年末，新西兰怀卡托大学计算机科学系Ian Written博士申请基金，1993年获新西兰政府资助，并于同年开发出接口和基础架构。1994年发布了第一个Weka的内部版本。1996年，第一个公开版本Weka 2.1发布。Weka的早期版本主要采用C语言编写。1997年，团队决定使用Java重新改写。Java允许用一个统一的接口来学习方案和方法的预处理和后处理，决定使用Java来代替C++或其他面向对象的语言，是因为Java编写的程序可以运行在绝大部分计算机上，而无需重新编译，更不要修改源代码。Weka已经测试过的平台包括Linux，Windows和Macintosh操作系统，甚至PDA。最后的可执行程序复制过来即可运行，完全绿色，不需求复杂安装。当然，Java也有其缺点，最大的问题是他在速度上有缺陷，执行一个Java程序比相应的C语言程序马上好几倍。1999年，发布Java的Weka3版本。2013年，Weka的最新版本是3.7.8，直至今日，最新版本是3.8.6

2.2 处理方法

Weka的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），同时Weka也是新西兰的一种鸟名，而Weka的主要开发者来自新西兰。Weka作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的[机器学习算法](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0%E7%AE%97%E6%B3%95/18635836" \t "_blank)，包括对数据进行预处理，分类，回归、聚类、[关联规则](https://baike.baidu.com/item/%E5%85%B3%E8%81%94%E8%A7%84%E5%88%99" \t "_blank)以及在新的交互式界面上的可视化。

2.2.1 分类与回归

Weka把分类(Classification)和回归(Regression)都放在"Classify"选项卡中，这是有原因的。

在这两个任务中，都有一个目标属性(输出变量)。我们希望根据一个样本(WEKA中称作实例)的一组特征(输入变量)，对目标进行预测。为了实现这一目的，我们需要有一个训练数据集，这个数据集中每个实例的输入和输出都是已知的。观察训练集中的实例，可以建立起预测的模型。有了这个模型，我们就可以新的输出未知的实例进行预测了。衡量模型的好坏就在于预测的准确程度。

在WEKA中，待预测的目标(输出)被称作Class属性，这应该是来自分类任务的"类"。一般的，若Class属性是分类型时我们的任务才叫分类，Class属性是数值型时我们的任务叫回归。

2.2.2 聚类

聚类分析中的"类"(cluster)和前面分类的"类"(class)是不同的，对cluster更加准确的翻译应该是"簇"。聚类的任务是把所有的实例分配到若干的簇，使得同一个簇的实例聚集在一个簇中心的周围，它们之间距离的比较近;而不同簇实例之间的距离比较远。对于由数值型属性刻画的实例来说，这个距离通常指欧氏距离。

2.2.3 关联规则

关联规则又称购物栏分析。

目前，Weka的[关联规则](https://baike.so.com/doc/6387919-6601574.html" \t "https://baike.so.com/doc/_blank)分析功能仅能用来作示范，不适合用来挖掘大型数据集。

我们打算对前面的"bank-data"数据作关联规则的分析。用"Explorer"打开"bank-data-final.arff"后，切换到"Associate"选项卡。默认关联规则分析是用Apriori算法，我们就用这个算法，但是点"Choose"右边的文本框修改默认的参数，弹出的窗口中点"More"可以看到各参数的说明。

背景知识

首先我们来温习一下Apriori的有关知识。对于一条关联规则L->R，我们常用支持度(Support)和置信度(Confidence)来衡量它的重要性。规则的支持度是用来估计在一个购物栏中同时观察到L和R的概率P(L,R)，而规则的置信度是估计购物栏中出现了L时也出会现R的条件概率P(R|L)。[关联规则](https://baike.so.com/doc/6387919-6601574.html" \t "https://baike.so.com/doc/_blank)的目标一般是产生支持度和置信度都较高的规则[1]。

2.3 输入数据

通过以ARFF格式为代表的文件进行输入，然后直接读取数据库表。

2.4 Weka主界面Weka GUI Chooser

Weka主界面为Weka GUI选择器，它通过右边的四个按钮提供四种主要的应用程序供用户选择。

（1）探索者（Explorer），系统提供的最容易使用的图像用户接口。通过选择菜单和填写表单，可以调用Weka的所有功能，最容易使用的图形用户界面，通过菜单和填写表单可以调用Weka的所有功能。

虽然探索者界面使用很方便，但它也存在一个缺陷，要求它将所需数据全部一次读进内存，一旦用户打开某个数据集，就会读取全部数据。因此，这种批量方式仅适合处理中小规模的问题。而知识流刚好能够弥补这一缺陷。

（2）知识流（KnowledgeFlow），可以使用增量方式的算法来处理大规模的数据集，无需一次性全部导入。使用增量方式的算法来处理大型数据集，用户可以定制处理数据流的方式和顺序。按照一定顺序将代表数据源、预处理工具、学习算法、评估手段和可视化模块的各构件组合在一起，形成数据流。

（3）实验者（Experimenter），用于帮助用户解答实际应用分类和回归技术中遇到的一个基本问题——对于一个已知问题，哪种方法及参数值能够取得最佳效果？尽管探索者界面也能通过交互完成这样的功能，但通过实验者界面，用户可以让处理过程实现自动化。实验者界面更加容易使用不同参数去设置分类器和过滤器，使之运行在一组数据集中，收集性能统计数据，实现重要的测试实验。

对于一个已知问题，哪种方法及参数值能够取得最佳效果。即可以用于进行多种算法性能的比较。ARFF格式是Weka的专用格式。

Weka整理了几门免费的在线课程，使用Weka教授机器学习和数据挖掘。课程视频可在Youtube上找到。

（4）简单命令行（Simple CLI），这个界面是为不提供自己的命令行界面的操作系统提供的，该简单命令行界面用户和用户进行交互，可以直接执行Weka命令[2]。

## 3 数据集合定义

3.1 数据格式

跟很多[电子表格](https://baike.baidu.com/item/%E7%94%B5%E5%AD%90%E8%A1%A8%E6%A0%BC" \t "https://baike.baidu.com/item/weka/_blank)或数据分析软件一样，Weka所处理的数据集是一个二维的表格。

这里我们要介绍一下Weka中的术语。表格里的一个横行称作一个实例（Instance），相当于统计学中的一个样本，或者数据库中的一条记录。竖行称作一个属性（Attribute），相当于统计学中的一个变量，或者数据库中的一个字段。这样一个表格，或者叫数据集，在Weka看来，呈现了属性之间的一种关系(Relation)。示例表格一共有14个实例，5个属性，关系名称为“weather”。

Weka存储数据的格式是ARFF（Attribute-Relation File Format）文件，这是一种ASCII文本文件。[二维表格](https://baike.baidu.com/item/%E4%BA%8C%E7%BB%B4%E8%A1%A8%E6%A0%BC" \t "https://baike.baidu.com/item/weka/_blank)存储在如下的ARFF文件中。这也就是Weka自带的“weather.arff” 文件，在Weka[安装目录](https://baike.baidu.com/item/%E5%AE%89%E8%A3%85%E7%9B%AE%E5%BD%95" \t "https://baike.baidu.com/item/weka/_blank)的”data”子目录下可以找到[1]。

Weka存储数据的格式是ARFF（Attribute-Relation File Format）文件，这是一种ASCII文本文件。[二维表格](https://baike.baidu.com/item/%E4%BA%8C%E7%BB%B4%E8%A1%A8%E6%A0%BC" \t "_blank)存储在如下的ARFF文件中。这也就是Weka自带的“weather.arff”文件，在Weka[安装目录](https://baike.baidu.com/item/%E5%AE%89%E8%A3%85%E7%9B%AE%E5%BD%95" \t "_blank)的“data”子目录下可以找到。

目前，Weka的[关联规则](https://baike.baidu.com/item/%E5%85%B3%E8%81%94%E8%A7%84%E5%88%99" \t "_blank)分析功能仅能用来作示范，不适合用来挖掘大型数据集。

表1-1 weather数据集合

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 索引值 | outlook | temperature | humidity | windy | play |
| 1 | sunny | hot | high | FALSE | no |
| 2 | sunny | hot | high | TRUE | no |
| 3 | overcast | hot | high | FALSE | yes |
| 4 | rainy | mild | high | FALSE | yes |
| 5 | rainy | cool | normal | FALSE | yes |
| 6 | rainy | cool | normal | TRUE | no |
| 7 | overcast | cool | normal | TRUE | yes |
| 8 | sunny | mild | high | FALSE | no |
| 9 | sunny | cool | normal | FALSE | yes |
| 10 | rainy | mild | normal | FALSE | yes |
| 11 | sunny | mild | normal | TRUE | yes |
| 12 | overcast | mild | high | TRUE | yes |
| 13 | overcast | hot | normal | FALSE | yes |
| 14 | rainy | mild | high | TRUE | no |

（1）weather.nominal.arrf介绍

@relation weather.symbolic命名表,表名为weather.symbolic（二维表）。

@attribute outlook命名列，列名为outlook。

@data命名行，实例具体的值。

### 3.2 术语

表格里的一个横行称作一个实例（Instance），相当于统计学中的一个样本，或者数据库中的一条记录。竖行称作一个属性（Attribute），相当于统计学中的一个变量，或者数据库中的一个字段。这样一个表格，或者叫数据集，在Weka看来，呈现了属性之间的一种关系（Relation）。

### 3.3 声明

声明分为关系声明和属性声明

#### 3.3.1 关系声明

关系名称在arff文件的第一个有效行来定义，格式为

@relation<relation-name>

<relation-name>是一个字符串。如果这个字符串包含空格，它必须加上引号（指英文标点的单引号或双引号）。

#### 3.3.2 属性声明

属性声明用一列以“@attribute”开头的语句表示。数据集中的每一个属性都有它对应的“@attribute”语句，来定义它的属性名称和数据类型。

这些声明语句的顺序很重要。首先它表明了该项属性在数据部分的位置。例如，“humidity”是第三个被声明的属性，这说明数据部分那些被逗号分开的列中，第三[列数据](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=74633135&ss_c=ssc.citiao.link" \t "https://baike.sogou.com/_blank)85 90 86 96 ...是相应的“humidity”值。其次，最后一个声明的属性被称作class属性，在分类或回归任务中，它是默认的目标变量。

属性声明的格式为

@attribute<attribute-name><datatype>

其中<attribute-name>是必须以字母开头的字符串。和关系名称一样，如果这个字符串包含空格，它必须加上引号。

Weka支持的<datatype>有四种，分别是

numeric-------------------------数值型

<nominal-specification>-----分类（nominal）型

string----------------------------字符串型

Date [<date-format>]--------日期和时间型

其中<nominal-specification>和<date-format>将在下面说明。还可以使用两个类型“integer”和“real”，但是Weka把它们都当作“numeric”看待。注意“integer”，“real”，“numeric”，“date”，“string”这些关键字是区分大小写的，而“relation”, “attribute”和“data”则不区分。

### 3.4 属性

属性大致分为五类

#### 3.4.1 数值属性

数值型属性可以是整数或者实数，但Weka把它们都当作实数看待。

#### 3.4.2 分类属性

分类属性由<nominal-specification>列出一系列可能的类别名称并放在花括号中：{<nominal-name1>,<nominal-name2>,<nominal-name3>,...}。数据集中该属性的值只能是其中一种类别。

例如如下的属性声明说明“outlook”属性有三种类别：“sunny”，“ overcast”和“rainy”。而数据集中每个实例对应的“outlook”值必是这三者之一。

@attribute outlook{sunny,overcast,rainy}

如果类别名称带有空格，仍需要将之放入引号中。

#### 3.4.3 字符串属性

字符串属性中可以包含任意的文本。这种类型的属性在文本挖掘中非常有用。

示例：

@ATTRIBUTE LCC string

#### 3.4.4 日期和时间属性

日期和时间属性统一用“date”类型表示，它的格式是

@attribute<name>date[<date-format>]

其中<name>是这个属性的名称，<date-format>是一个字符串，来规定该怎样解析和显示日期或时间的格式，默认的字符串是ISO-8601所给的日期时间组合格式“yyyy-MM-ddTHH:mm:ss”。

数据信息部分表达日期的字符串必须符合声明中规定的格式要求。

#### 3.4.5 Relational型属性

在Weka3.5版中增加了一种属性类型叫做Relational，有了这种类型我们可以像[关系型数据库](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=277136&ss_c=ssc.citiao.link" \t "https://baike.sogou.com/_blank)那样处理多个维度了。但是这种类型目前还不见广泛应用，暂不作介绍。

### 3.5 数据

一般的数据分为数据信息和稀疏数据

#### 3.5.1 数据信息

数据信息中“@data”标记独占一行，剩下的是各个实例的数据。

每个实例占一行。实例的各属性值用逗号“,”隔开。如果某个属性的值是缺失值（missing value），用问号“?”表示，且这个问号不能省略。例如：

@data

sunny,85,85,FALSE,no

?,78,90,?,yes

字符串属性和分类属性的值是区分大小写的。若值中含有空格，必须被引号括起来。例如：

@relation LCCvsLCSH

@attribute LCC string

@attribute LCSH string

@data

AG5, 'Encyclopedias and dictionaries.;Twentieth century.'

AS262, 'Science -- Soviet Union -- History.'

日期属性的值必须与属性声明中给定的相一致。例如：

@RELATION Timestamps

@ATTRIBUTE timestamp DATE "yyyy-MM-dd HH:mm:ss"

@DATA

"2001-04-03 12:12:12"

"2001-05-03 12:59:55"

#### 3.5.2 稀疏数据

有的时候数据集中含有大量的0值（比如购物篮分析），这个时候用稀疏格式的数据存贮更加省空间。

稀疏格式是针对数据信息中某个实例的表示而言，不需要修改arff文件的其它部分。看如下的数据：

@data

0, X, 0, Y, "class A"

0, 0, W, 0, "class B"

用稀疏格式表达的话就是

@data

{1 X, 3 Y, 4 "class A"}

{2 W, 4 "class B"}

每个实例用花括号括起来。实例中每一个非0的属性值用<index> <空格> <value>表示。<index>是属性的序号，从0开始计；<value>是属性值。属性值之间仍用逗号隔开。这里每个实例的数值必须按属性的顺序来写，如{1 X, 3 Y, 4 "class A"}，不能写成{3 Y, 1 X, 4 "class A"}。

注意在稀疏格式中没有注明的属性值不是缺失值，而是0值。若要表示缺失值必须显式的用问号表示出来。

### 3.6 csv转换为arff格式

我们通常接触到的数据为csv格式较多，若想要将数据更改为Weka通用的arff格式，我们可以利用Weka工具进行格式的转换。

#### 3.6.1 csv格式

逗号分隔值（Comma-Separated Values，CSV，有时也称为字符分隔值，因为分隔字符也可以不是逗号），其文件以纯文本形式存储表格数据（数字和文本）。纯文本意味着该文件是一个字符序列，不含必须像二进制数字那样被解读的数据。CSV文件由任意数目的记录组成，记录间以某种换行符分隔；每条记录由字段组成，字段间的分隔符是其它字符或字符串，最常见的是逗号或制表符。通常，所有记录都有完全相同的字段序列。通常都是纯文本文件。建议使用Word或是记事本来开启，再则先另存新档后用Excel开启，也是方法之一[4]。

#### 3.6.2 arff格式

arff是一种Weka专用的文件格式，由Andrew Donkin创立，有传言说arff代表Andrew's Ridiculous File Format(安德鲁的荒唐文件格式)，但在Weka的正式文档中明确说明arff代表Attribute-Relation File Format(属性——关系文件格式)。该文件是ASCII文本文件，描述共享一组属性结构的实例列表，由独立且无序的实例组成，是Weka表示数据集的标准方法，arff不涉及实例之间的关系。

#### 3.6.3 获取csv文件

本次案例我们使用UCI网站中的audiology数据进行演示，我们先从网站中下载下来我们所需要的数据，进入网站后选择Download按钮进入下载页面，我们选择audiology.data数据进行下载下来发现是一个纯文本文件，用记事本打开，发现此文件大约有200多条实例，属性多达36个，为了方便用表展示，我们只截取了前5个，实例也仅选取了前9个以作示范。

将网站主页中的属性名依次复制到该data文件中，属性之间用逗号隔开，这样我们就构造出来了csv格式的数据文件。保存文件，将该文件的后缀改为csv，我们就得到这个数据集的csv格式文件。

表3-1 audiology数据集合

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | class | acoustic\_neuroma | bells\_palsy | cochlear\_age | cochlear\_age\_and\_noise |
| 1 | cochlear\_unknown | boneAbnormal | air(mild) | ar\_c(normal) | ar\_u(normal) |
| 2 | cochlear\_unknown | boneAbnormal | air(moderate) | ar\_c(normal) | ar\_u(normal) |
| 3 | mixed\_cochlear\_age\_fixation | age\_gt\_60 | airBoneGap | boneAbnormal | air(mild) |
| 4 | mixed\_cochlear\_age\_otitis\_media | age\_gt\_60 | airBoneGap | air(mild) | ar\_u(absent) |
| 5 | cochlear\_age | age\_gt\_60 | boneAbnormal | air(mild) | ar\_c(normal) |
| 6 | cochlear\_age | age\_gt\_60 | boneAbnormal | air(mild) | ar\_c(normal) |
| 7 | cochlear\_unknown | boneAbnormal | air(mild) | ar\_c(normal) | ar\_u(normal) |
| 8 | cochlear\_unknown | boneAbnormal | air(mild) | ar\_c(normal) | ar\_u(normal) |
| 9 | cochlear\_unknown | boneAbnormal | air(severe) | static(normal) | tymp(a) |

## 4 广告点击率预估挑战赛的预测与分析

广告点击率预估是在线广告交易的核心环节之一，如果说一家公司胡知道CTR(点击率)，以确定将他们的钱花在数字广告上是否值得。点击率高表示对孩核特定广告系列更感兴趣，点击率低可能表明广告可能不那么相关。高点击率表明更多人点击了网站，这有利于在谷歌、必应等在线平台上以更少的钱获得更好的。

近年来，各大有关广告点击率预估的比赛相拥而至，如腾讯广告算法大赛、科大讯飞营销算法大赛、阿里点击率预大赛等。可以看出这是一个企业长期关注的问题，也是值得花时间探索的问题。

### 4.1 数据理解

#### 4.1.1 数据集示例

广告点击率预估赛数据集合来自于科大讯飞2021广告点击率预估挑战赛数据，网址:http:/lchallenge.xfyun.cn/topic/info?type=Ad-click-through。包含训练集和测试集。其中测试集一共提供了71,466个样本，12个特征字段;训练集一共提供了391,825个样本和13个特征。由于数据样本庞大，我只截取测试集中6个特征，和9个数据。

表4-1 test数据集合

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| id | date | user\_id | product | campaign\_id | webpage\_id |
| 391825 | 2022/7/7 0:00 | 94025 | 0 | 2 | 2 |
| 391826 | 2022/7/7 0:00 | 135659 | 0 | 2 | 2 |
| 391827 | 2022/7/7 0:00 | 135659 | 4 | 9 | 8 |
| 391828 | 2022/7/7 0:00 | 2569 | 3 | 9 | 8 |
| 391829 | 2022/7/7 0:00 | 53895 | 7 | 4 | 3 |
| 391830 | 2022/7/7 0:00 | 81998 | 9 | 9 | 8 |
| 391831 | 2022/7/7 0:01 | 17921 | 3 | 4 | 3 |
| 391832 | 2022/7/7 0:01 | 44038 | 6 | 9 | 8 |
| 391833 | 2022/7/7 0:01 | 81817 | 3 | 0 | 0 |

#### 4.1.1 数据集属性介绍

广告点击率预估数据集合的测试集里共有13个，分别是id,date,user\_id,product,campaign\_id,webpage\_id,product\_category\_id,user\_group\_id,gender,age\_level,user\_depth,var\_1,isClick。他们的类型分为两种，分别是Numeric和Nominal。

表4-2 字段取值数量

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段取值数量 | 类型 |
| id | 391825 | Numeric |
| date | 7174 | Nominal |
| user\_id | 135658 | Numeric |
| product | 10 | Numeric |
| campaign\_id | 10 | Numeric |
| webpage\_id | 9 | Numeric |
| product\_category\_id | 5 | Numeric |
| user\_group\_id | 13 | Numeric |
| gender | 2 | Nominal |
| age\_level | 7 | Numeric |
| user\_depth | 3 | Numeric |
| var\_1 | 2 | Numeric |
| isClick | 2 | Numeric |

### 4.2 实验过程及结果分析

本文的日的是为了研究在点击率预估中，对字段进行分布表示后在模型性能上带来的提升。因此，本文首先进行了模型单独的训练，然后添加了基于奇异值分解的向量重新训练，最后再使用Skip-gram模型训练并对结果进行比较。

#### 4.2.1 模型训练

模型方面使用了三种:梯度提升树的集成学习模型（GBDT)、域感知因子分解机（FFM）模型和深度因子分解机（DeepFM）模型。LightGBM是微软提供的基于GBDT的模型，在XGBoost的基础上优化了决策树的分裂算法和生长策略，训练效率上有比较大的提升，并且采用了互斥特征绑定，能直接支持类别标签特征，不需要独热编码，减少了内存占用。FFM使用了libFFM包训练，libFMM是由阮毓钦团队提供的域感知因子分解机的软件包。深度因子分解机模型则参考了论文和部分开源实现，使用了Tensorflow进行复现。

（1）梯度提升树

首先是单独使用LightGBM包训练。由于LightGBM直接支持类别标签特征，因此基于上一节得到的数据可以不加处理地输入LightGBM中。在训练过程中，由于数据量较大，我们使用了贪心的思路，逐个超参数进行调参，调参过程使用交叉验证的方式进行。

1. 域感知因子分解机

其次是单独使用libFFM包进行训练。域感知因子分解机模型的输入比较特殊，需要指定特征所属的域和特征的编号，并转换为如下格式:

label,field,:indexy:value1,fieldz:indexz vatuez2,...

0,0:0:0.1,1:1:0.5,2:3:0.2,...

0,0:0:0.2,1:2:0.3,2:5:0.1....1,0:0:0.2,1:2:0.3,2:5:0.1,...

其中label为样本的标签，field为特征所属的域，index为特征编号，value为特征值。这种类型的输入格式是为了便于存放大规模的稀疏特征数据，运算时也仅使用输入中存在的特征，提高了计算效率。最自然的想法当然是每一个原始特征当作一个域，域内的每一个值当作当作一个稀疏特征。本文在上一节中分析了各个特征的分布特点，device\_id字段和 device\_ip字段的头部和非头部数据差异明显，取值数也达到了百万级别，由于尾部样本过于稀疏，这部分参数非常难以训练。因此，本文对device\_id字段和 device\_ip字段进行了截断,device\_id出现少于8次的值统一其中label为样本的标签，field为特征所属的域，index为特征编号，value为特征值。这种类型的输入格式是为了便于存放大规模的稀疏特征数据，运算时也仅使用输入中存在的特征，提高了计算效率。最自然的想法当然是每一个原始特征当作一个域，域内的每一个值当作当作一个稀疏特征。本文在上一节中分析了各个特征的分布特点，device\_id字段和 device\_ip字段的头部和非头部数据差异明显，取值数也达到了百万级别，由于尾部样本过于稀疏，这部分参数非常难以训练。因此，本文对device\_id字段和device\_ip字段进行了截断,device\_id出现少于8次的值统一。

### 4.3 模型对比

#### 4.3.1 逻辑回归模型

逻辑回归模型(Logistic regression，LR)，又称对数几率模型。由于逻辑回归模型简单，可解释强，易实现，广泛应用于机器学习、深度学习、推荐系统、广告预估、智能营销、金融风控、社会学、生物学、经济学等领域。现在看来，LR依然是推荐系统、广告预估的三大基础模型之一。如果想从事推荐相关、广告预估的工作、研究、学习，是必须要掌握的。

#### 4.3.2 随机森林模型

#### 4.3.3 LightGBM

### 4.4 特征选择

#### 4.4.1 arff格式

### 4.5 模型训练与评估

#### 4.5.1 arff格式

## 参考文献

1. 百度百科, URL: https://baike.baidu.com/item/weka/10701215?fr=aladdin. [2022-6-16].

[2] zuokankan, URl: http://t.zoukankan.com/chamie-p-4551514.html. [2022-6-16].

[3] CSDN, URL: <https://blog.csdn.net/qq_32892383/article/details/104424358.> [2022-6-17]