手机用户分类模型

摘 要

随着人们使用手机的频率越来越高，电信运营商掌握着大量的手机用户的使用数据。如何充分利用这些海量的用户数据，从中挖掘出有价值的知识，是当前电信运营商们面临的一个重要课题。有了这些知识，电信运营商可以采取更好的商业策略，也可以对行为异常的用户进行预测。本文使用数据挖掘和机器学习的方法，采取常见的分类和异常检测的算法对用户数据进行分析，预测异常用户，从而对他们采取相应的措施。

关键词：机器学习，大数据，数据挖掘，分类算法，异常检测

### Models for mobile users classification

# **Abstract**

As people use mobile phones more and more frequently, telecom operators hold a lot of data on the usage of mobile phone users. How to fully make use of the vast amount of subscriber data and extract valuable knowledge from it is an important topic facing telecom operators nowadays. With this knowledge, telecom operators can adopt better business strategies and also make predictions about users with abnormal behavior. This paper uses classification and anomaly detection algorithms of data mining and machine learning to analyze user data and then predict anomalous users so that action can be taken to against them.

**Keywords**：Data Mining；Machine Learning；Big Data;Classification algorithm; ano

目 录

1. 概述 （1）

1.1 研究背景 （1）

1.2 数据轮廓 （1）

2.数据预处理 （ ）

2.1 （ ）

2.1.1 （ ）

3. （ ）

3.1 （ ）

3.1.1 （ ）

4. （ ）

4.1 （ ）

4.1.1 （ ）

5. （ ）

5.1 （ ）

5.1.1 （ ）

参考文献 （ ）

# 1.概 述

## **1.1 研究背景**

近十年多来，随着中国手机用户的数量爆炸增长，中国的手机用户基数已经十分庞大。根据中国工信部的数据显示，2018年中国手机用户总数已达15.7亿。如此庞大的用户群体，必然产生海量的通信数据。在保障用户隐私的前提下，人们很早就有分析这些数据的想法，而数据挖掘和机器学习技术的进步让人们有了更好的工具去从海量的数据中发现有价值的知识。比如建立电信客户流失模型，实现精确营销，分析客户关系等等都是数据挖掘在电信领域的成功应用。

数据挖掘（Data Mining）是从大量的、不完全的、有噪音的、模糊的、随机的实际应用数据中，提取隐含在其中人们事先不知道的、但又是潜在有用的信息和知识的过程。它是一个跨学科的计算机科学分支，与人工智能、机器学习、统计学和数据库等领域密不可分。

机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。

本文的数据集为60万条已脱敏的中国移动公司的客户数据，其中有500条是被投诉的用户的信息。本文的目的是，利用机器学习的分类算法和异常检测模型，预测未来可能会被投诉的用户。

## **1.2 数据轮廓**

原始数据集中有600023位客户的数据，其中502位客户被投诉。数据有27维，其中6维是布尔型数据，名称分别为：'users\_3w', 'twolow\_users', 'roam\_users02', 'roam\_users01','vv\_type', 'in16\_roam\_tag'，其余21维为数值型数据，名称分别为'roam\_call\_duration','roam\_duration\_02','mon\_use\_days','is\_p\_app\_wx\_times', 'zhujiao\_time','zhujiao\_times','mb5','mb10','mb30','mb60','ma60','total\_count','beijiao\_times', 'use\_days', 'zhujiao', 'beijiao','zhujiao\_jt', 'open', 'close', 'open\_day', 'cell\_num'。

# 数据预处理

数据的来源有很多，网上电子银行的操作会产生数据，人们使用手机会产生数据，问卷调查也会产生数据，等等。但是这些数据往往是不干净、不完全、有噪音、不一致、冗余的。这些数据还不能被直接分析，我们需要对数据做很多预处理工作。数据预处理是数据分析的基础。

## **数据清洗**

* + 1. **缺失数据**

处理缺失数据是数据清洗的一项重要工作。数据缺失的原因有很多，可能是因为收集数据的机器损坏了，也可能是因为涉及数据提供者的隐私。

处理数据缺失的方法包括但不限于以下3种：

1. 忽视

把含有缺失数据的属性去除。但是如果缺失的数据太多，这种方式会导致可用的数据很少。

1. 手工填写

数据收集者可以重新收集缺少的数据，但是在一些情况下，人们往往无法重新收集缺失的数据，而且这种方法也会增加成本。

另外一种方式是根据领域知识去猜测缺失的数据。比如一个人的性别为男，年龄是19岁，那么可以推测出他的体重为60公斤左右。

但是手工填写缺失数据是一个枯燥繁琐的过程。

1. 自动填写

借助计算机的帮助自动填写缺失数据。比如在缺失的地方填个常数或者该属性的平均值、中位数、均值。

在该数据集中，各个属性的频次直方图类似图2.1.1。离群点很突出，采用中位数来替换缺失值。

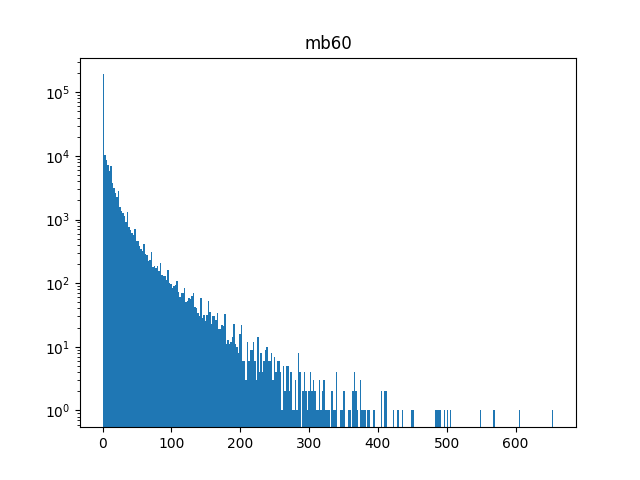


图2.1.1

* + 1. **离群点检测**

离群点是指不符合一般模型的数据对象，它们与数据的其它部分不同。如果不处理这些值，算法模型的结果可能会有很大的误差。

检测离群点的方法大致可分为一下3类：

1. 基于统计的离群点探测

假设数据符合某一个分布或者概率模型，然后根据模型采用不一致性检验：验证一个对象O关于分布F是否有显著不同。例如正态分布的3σ检测。

1. 基于距离的离群点探测

不依赖统计检验，基于距离的离群点探测将离群点看作是没有足够邻居的对象。如果一个数据点在其局部邻域内的点少与某个阈值，那么认为此数据点为异常点。

在这些方法中，最具有代表性的方法是局部离群因子检测方法 (Local Outlier Factor, LOF)。

1. 基于偏离的离群点探测

不同于基于统计和距离的离群点探测，基于偏离的离群点探测通过检查一组对象的主要特征来确定孤立点 ，如果与给出的描述偏离大的对象被认为是孤立点 。

观察本数据集某特征的分位数图（图2.1.2-1），我们发现，数据存在明显的离群点，而经过检查，这些数据点代表的用户并非被投诉的用户，所以我们用中位数代替这些离群点。

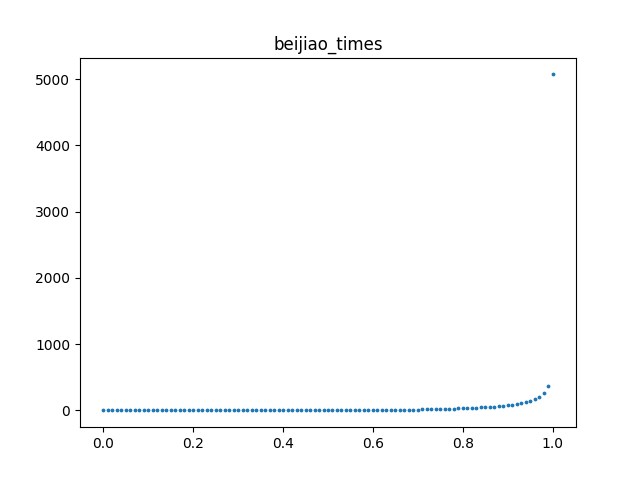


图2.1.2-1

对于本数据集，我们使用IQR（InterQuartile Range）和LOF算法去除离群点。

计算第一和第三个四分位数（Q1、Q3），定义,然后将大于Q3+3IQR和小于Q1-3IQR的数定义为离群值。

观察图2.1.1,发现很多正常数据的值很小，于是我们修改该方法，将大于Q3+3IQR数定义为离群值。

最后我们使用LOF算法，去掉邻居较少的数据点。

* 1. 特征选择
  2. 特征提取
  3. 数据归一化

参考文献：

[1] 王珊．数据库系统概论[M]．北京：高等教育出版社，2000．302-307．