个推 GeTui

《浅谈大数据与AI结合的一些思路和应用》

令狐冲

2020.12.26

目录

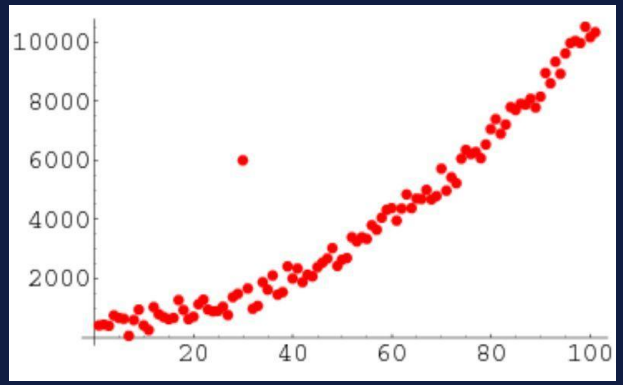
个推 GeTui

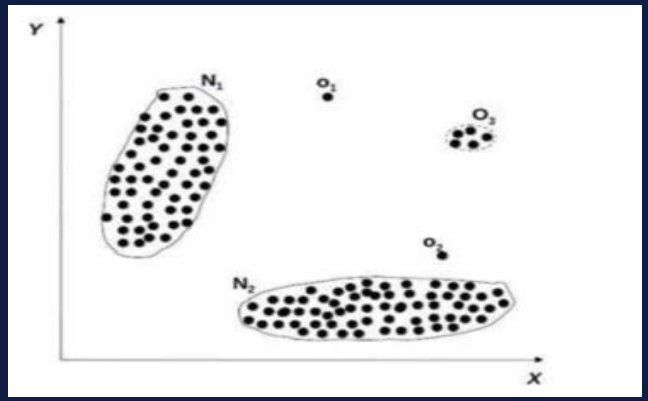
研究背景

个推 GeTui

研究背景

异常值（业务描述） ⟺ 离群点（数学描述）





应用场景：

1、流量监控

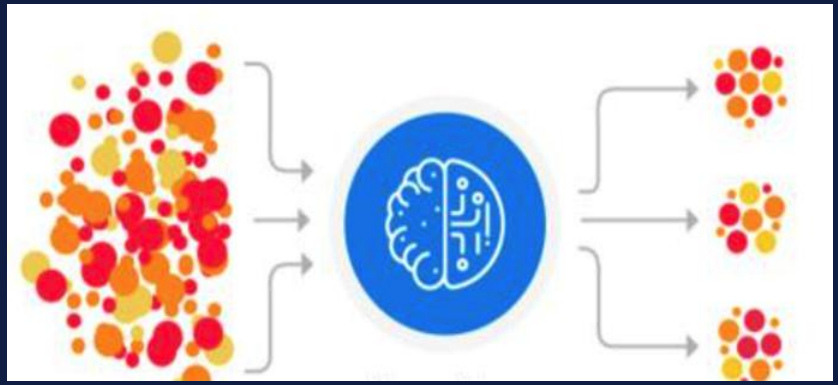
2、金融欺诈

3、系统故障检测

异常值检测方法

异常值检测方法

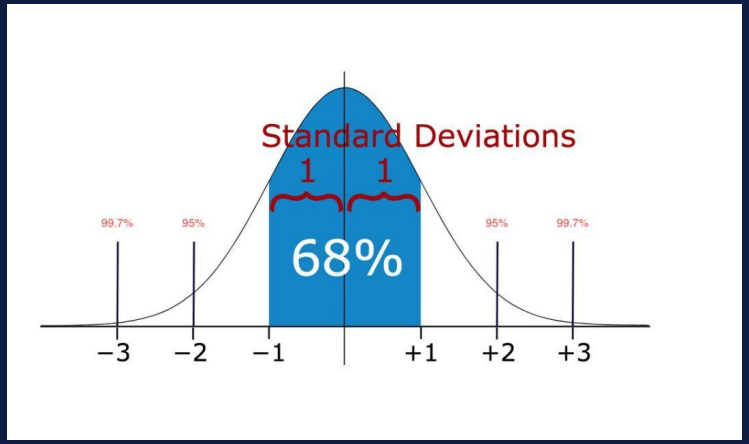
基于统计的方法



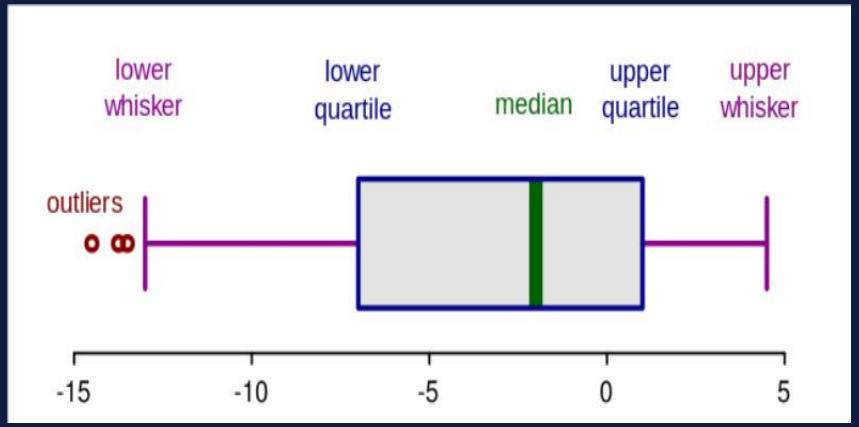
基于模型的方法

基于统计的异常值检测

基于统计的异常值检测

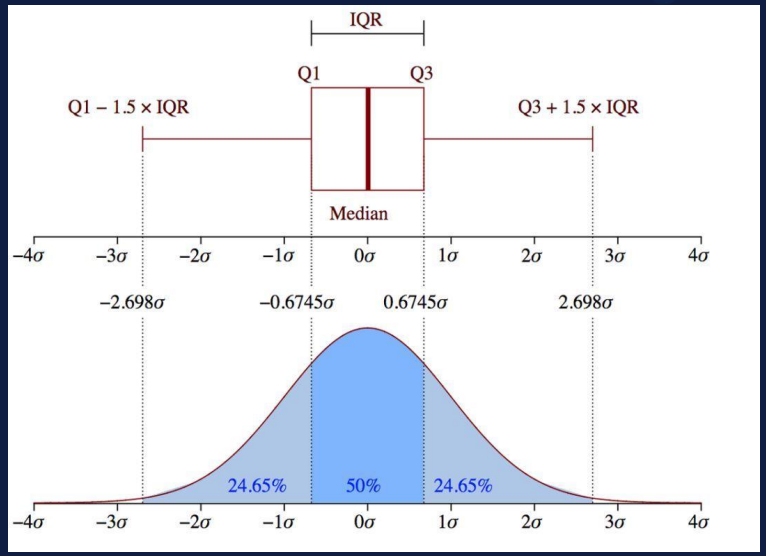


3σ法则



箱体图

基于统计的异常值检测



优点：

1）简单方便

2）坚实的统计学基础

缺点：

1）需要较多的样本数据

2）高维数据较难处理

基于模型的异常值检测

基于模型的异常值检测

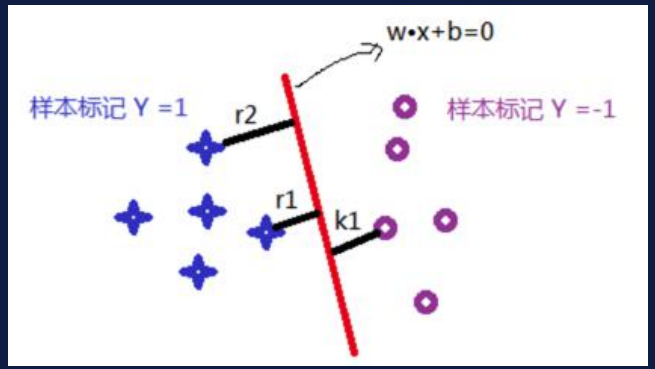
数学转化：分类问题，样本不均衡

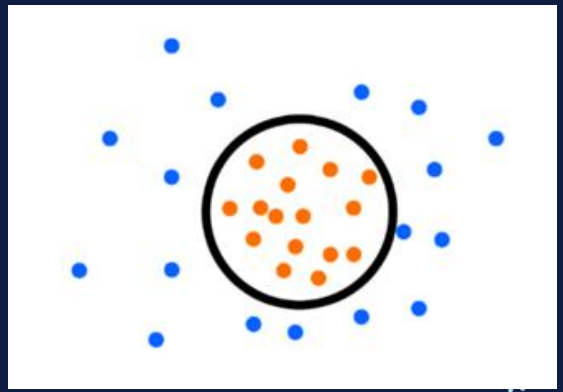
一、有监督模型：

k-最近邻：KNN

支持向量机：SVM（OneclassSVM）

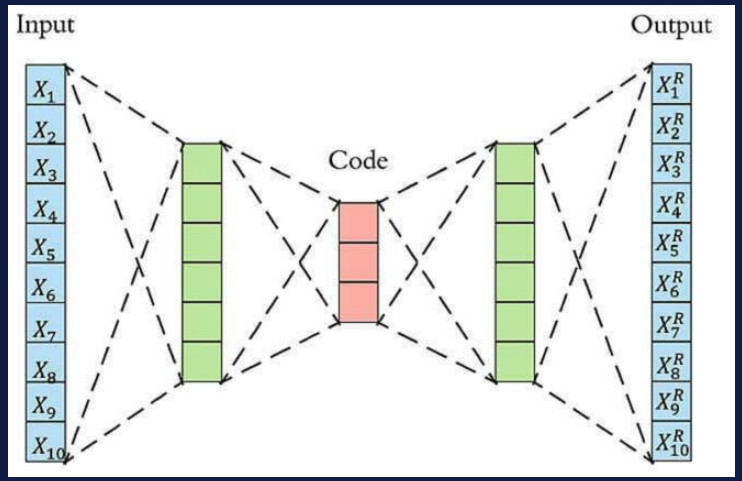
神经网络：自编码器（VAE）



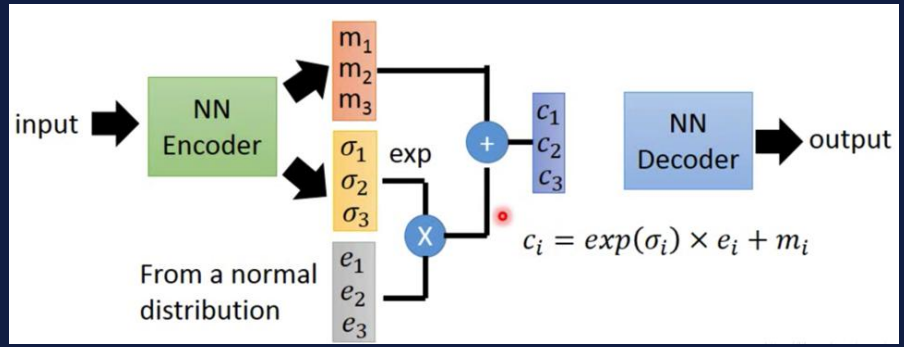


基于模型的异常值检测

神经网络方法



自编码器（AE）



变分自编码器（VAE）

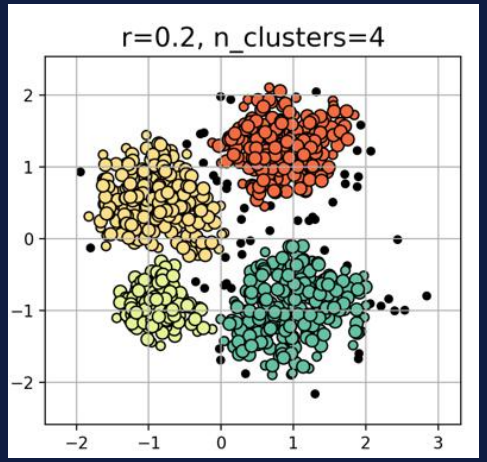
基于模型的异常值检测

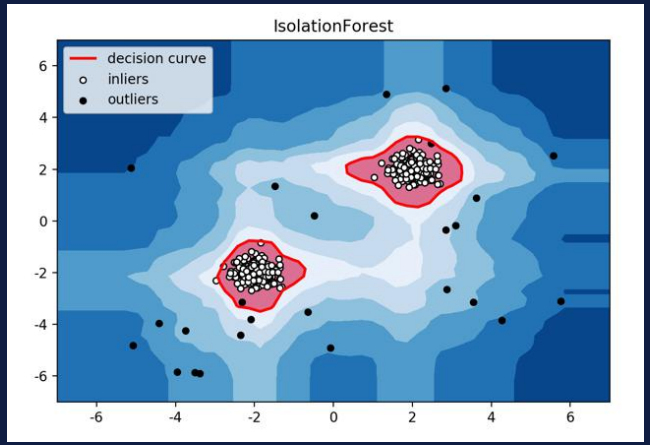
二、无监督模型：

密度聚类：DBSCAN

孤立森林：IsolationForest（IF）

随机剪切森林：RadomCutForest（RCF）

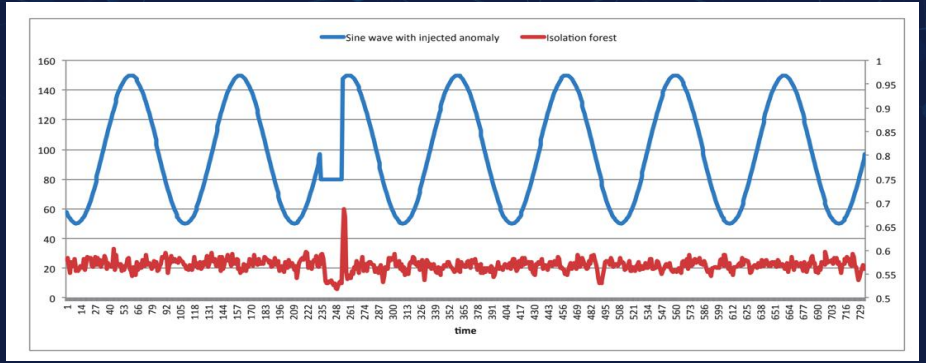




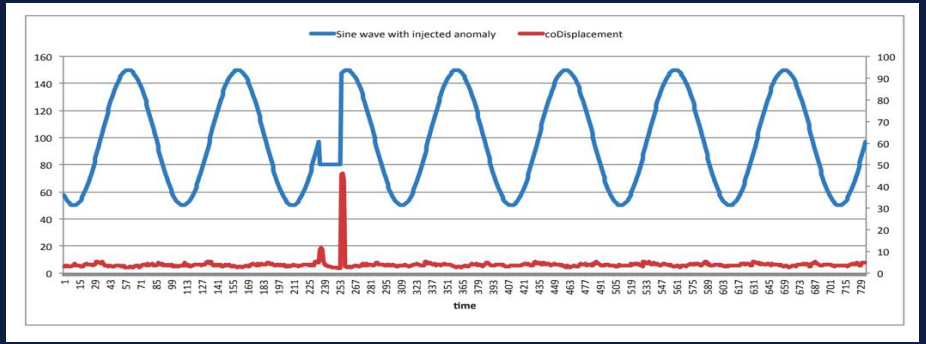
基于模型的异常值检测

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}$$

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$



IF算法

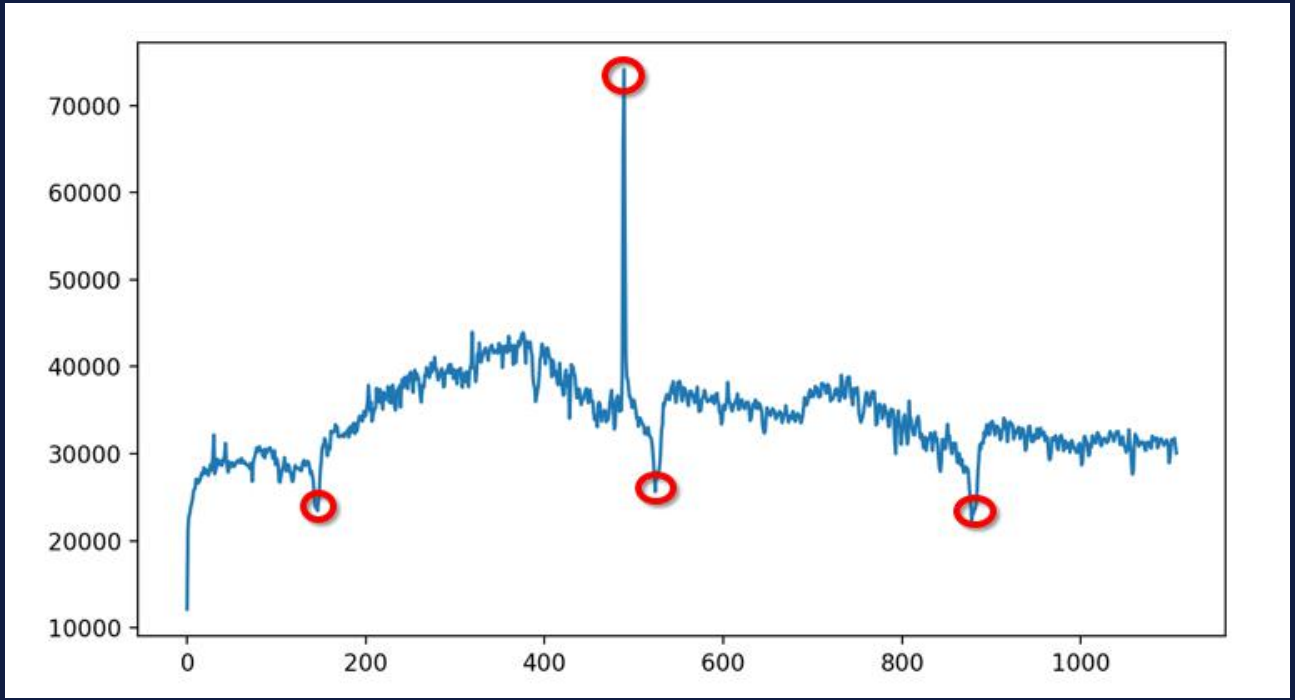


RCF算法

异常值检测实战应用

异常值检测实战应用

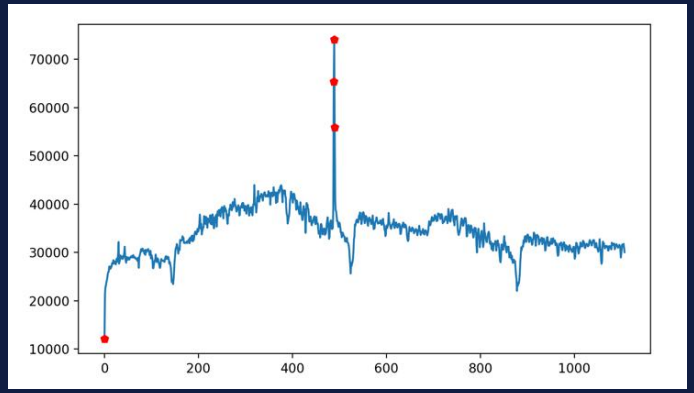
某APP从2016.08.16至2019.09.21的日活如下图所示：



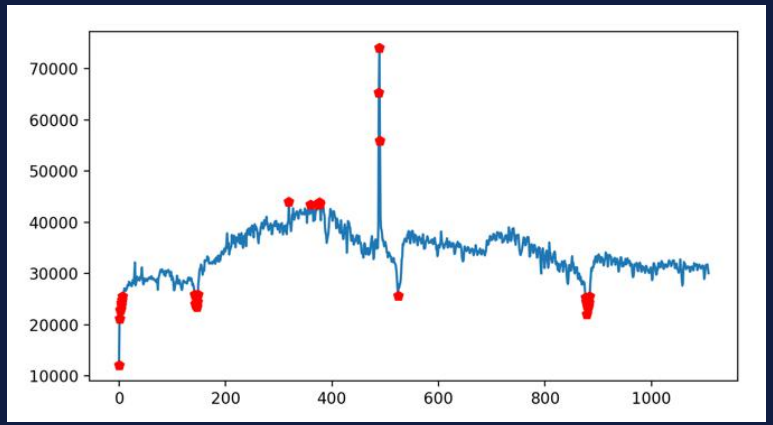
个推 GeTui

异常值检测实战应用

当成静止数据时的异常值检测：



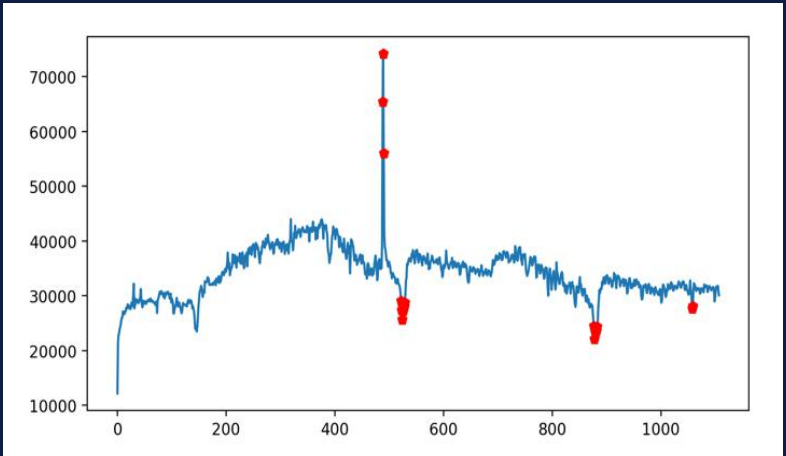
基于统计的3σ法则



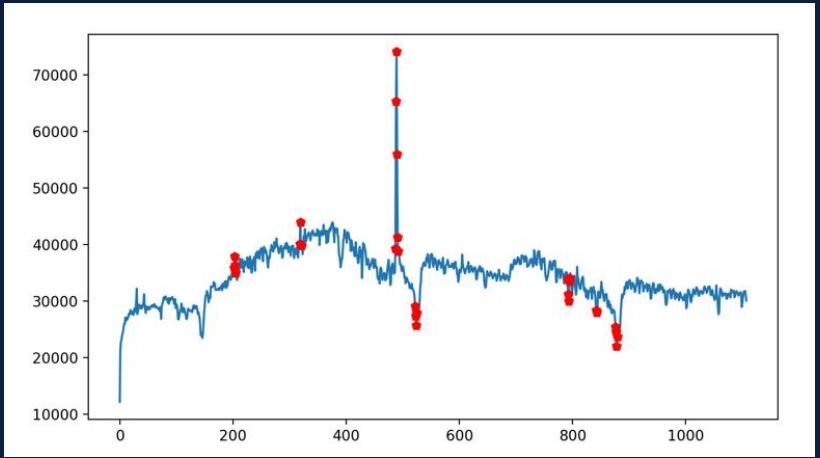
孤立森林算法

异常值检测实战应用

当成流式数据时异常值检测：

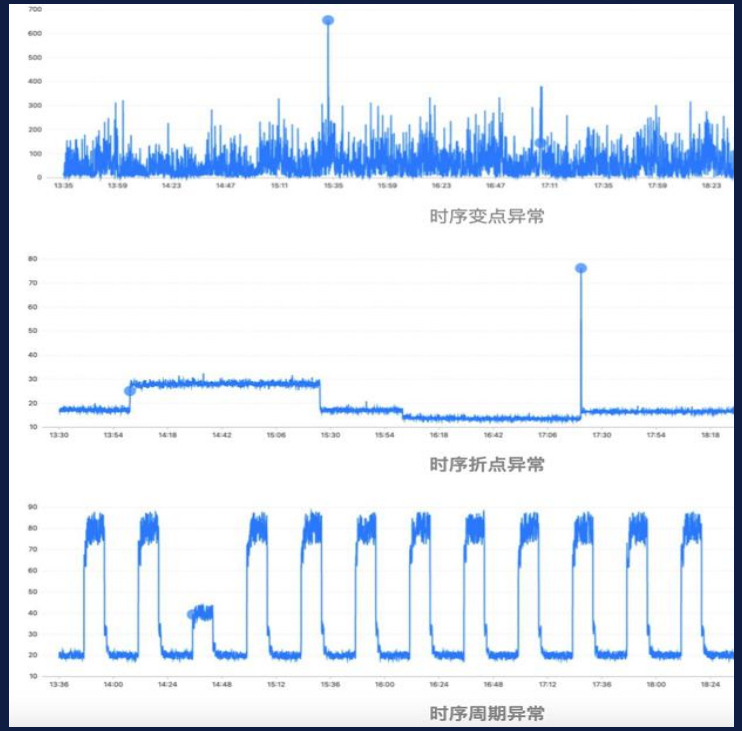


基于统计的滑动3σ法则



RCF算法

异常值检测实战应用



	变点异常	折点异常	周期异常
流式统计算法异常检测	强	中	弱
流式树算法异常检测	中	强	中
流式图算法异常检测	强	中	强

结合实际业务，合适才是最好

个推 GeTui

公司简介

个推成立于2010年，是专业的数据智能服务商，致力于用数据让产业更智能。公司以海量的数据积累和创新的技术理念，构建了移动开发、用户增长、品牌营销、公共管理和智能风控等多领域的数据智能服务生态。

个推十周年感恩特惠月

活动截止至：2020年12月31日

爆款开发者工具

消息推送VIP	免费用一年
用户画像	免费用一年
应用统计	免费用一年

热门大数据产品

人群洞察工具	免费试用
人口数盘	免费试用

立即扫码领取福利

个推 GeTui

个推公众号

个推技术学院公众号

数据让产业更智能

EMPOWER WITH DATA

个推 GeTui