一、概述

1、什么是异常检测

异常检测（Outlier Detection），顾名思义，是识别与正常数据不同的数据，与预期行为差异大的数

据。

识别如信用卡欺诈，工业生产异常，网络流里的异常（网络侵入）等问题，针对的是少数的事件。

1.1 异常的类别

点异常：指的是少数个体实例是异常的，大多数个体实例是正常的，例如正常人与病人的健康指标；

上下文异常：又称上下文异常，指的是在特定情境下个体实例是异常的，在其他情境下都是正常的，例

如在特定时间下的温度突然上升或下降，在特定场景中的快速信用卡交易；

群体异常：指的是在群体集合中的个体实例出现异常的情况，而该个体实例自身可能不是异常，例如社

交网络中虚假账号形成的集合作为群体异常子集，但子集中的个体节点可能与真实账号一样正常。

1.2 异常检测任务分类

有监督：训练集的正例和反例均有标签

无监督：训练集无标签

半监督：在训练集中只有单一类别（正常实例）的实例，没有异常实例参与训练

1.3 异常检测场景

故障检测

物联网异常检测

欺诈检测

工业异常检测

时间序列异常检测

视频异常检测

日志异常检测

医疗日常检测

网络入侵检测

2、异常检测常用方法

2.1 传统方法

2.1.1 基于统计学的方法

统计学方法对数据的正常性做出假定。**它们假定正常的数据对象由一个统计模型产生，而不遵守该模型**

的数据是异常点。统计学方法的有效性高度依赖于对给定数据所做的统计模型假定是否成立。

异常检测的统计学方法的一般思想是：学习一个拟合给定数据集的生成模型，然后识别该模型低概率区

域中的对象，把它们作为异常点。

即利用统计学方法建立一个模型，然后考虑对象有多大可能符合该模型。

假定输入数据集为，数据集中的样本服从正态分布，即，我们可

以根据样本求出参数和。

2.1.2 线性模型

典型的如PCA方法，Principle Component Analysis是主成分分析，简称PCA。它的应用场景是对数据集

进行降维。降维后的数据能够最大程度地保留原始数据的特征（以数据协方差为衡量标准）。

PCA的原理是通过构造一个新的特征空间，把原数据映射到这个新的低维空间里。PCA可以提高数据的

计算性能，并且缓解"高维灾难"。

2.1.3 基于相似度的方法

这类算法适用于数据点的聚集程度高、离群点较少的情况。同时，因为相似度算法通常需要对每一

个数据分别进行相应计算，所以这类算法通常计算量大，不太适用于数据量大、维度高的数据。

基于相似度的检测方法大致可以分为三类：

基于集群（簇）的检测，如DBSCAN等聚类算法。

聚类算法是将数据点划分为一个个相对密集的“簇”，而那些不能被归为某个簇的点，则被视作

离群点。这类算法对簇个数的选择高度敏感，数量选择不当可能造成较多正常值被划为离群点或成

小簇的离群点被归为正常。因此对于每一个数据集需要设置特定的参数，才可以保证聚类的效果，

在数据集之间的通用性较差。聚类的主要目的通常是为了寻找成簇的数据，而将异常值和噪声一同

作为无价值的数据而忽略或丢弃，在专门的异常点检测中使用较少。

聚类算法的优缺点：

（1）能够较好发现小簇的异常；

（2）通常用于簇的发现，而对异常值采取丢弃处理，对异常值的处理不够友好；

（3）产生的离群点集和它们的得分可能非常依赖所用的簇的个数和数据中离群点的存在性；

（4）聚类算法产生的簇的质量对该算法产生的离群点的质量影响非常大。

基于距离的度量，如k近邻算法。

k近邻算法的基本思路是对每一个点，计算其与最近k个相邻点的距离，通过距离的大小来判断它是否为离群点。在这里，离群距离大小对k的取值高度敏感。如果k太小（例如1），则少量的邻

近离群点可能导致较低的离群点得分；如果k太大，则点数少于k的簇中所有的对象可能都成了离群

点。为了使模型更加稳定，距离值的计算通常使用k个最近邻的平均距离。

k近邻算法的优缺点：

（1）简单；

（2）基于邻近度的方法需要O(m2)时间，大数据集不适用；

（3）对参数的选择敏感；

（4）不能处理具有不同密度区域的数据集，因为它使用全局阈值，不能考虑这种密度的变化。

基于密度的度量，如LOF（局部离群因子）算法。

局部离群因子（LOF）算法与k近邻类似，不同的是它以相对于其邻居的局部密度偏差而不是距离来进行度量。它将相邻点之间的距离进一步转化为“邻域”，从而得到邻域中点的数量

（即密

度），认为密度远低于其邻居的样本为异常值。

LOF（局部离群因子）算法的优缺点：

（1）给出了对离群度的定量度量；

（2）能够很好地处理不同密度区域的数据；

（3）对参数的选择敏感。__