**数据集**

**SMD**

服务器机器数据集(Server Machine Dataset)

SMD是从一家大型互联网公司获得的数据集，收集28个不同机器的数据，总的时间周期为五周。数据集被均等分为训练集和测试集，并且各包含28个实体，每个实体包含了38个维度，其中训练集数据有708,405个样本，测试集有708,420个样本（1416825），训练数据集中的异常由专家标记，最后选择28个实体的平均分数作为最终的异常得分。

28个机器五周的数据，时间间隔1min，（五周总共有50400分钟，意味着每台机器要检测50401次，总共28台机器，则总共检测50401×28 = 1411228次）每个机器有38个特征。训练集和测试集的数据比例为1:1，适用于无监督异常检测模型。**用SMD数据集跑模型的时候，28个机器的数据需要分开训练。**

**2.原文件**

28个机器分成了三组。（第一组有8个，第二组有9个，第三组10个）文件命名：machine-x-y.txt。第x组的第y个机器的特征。

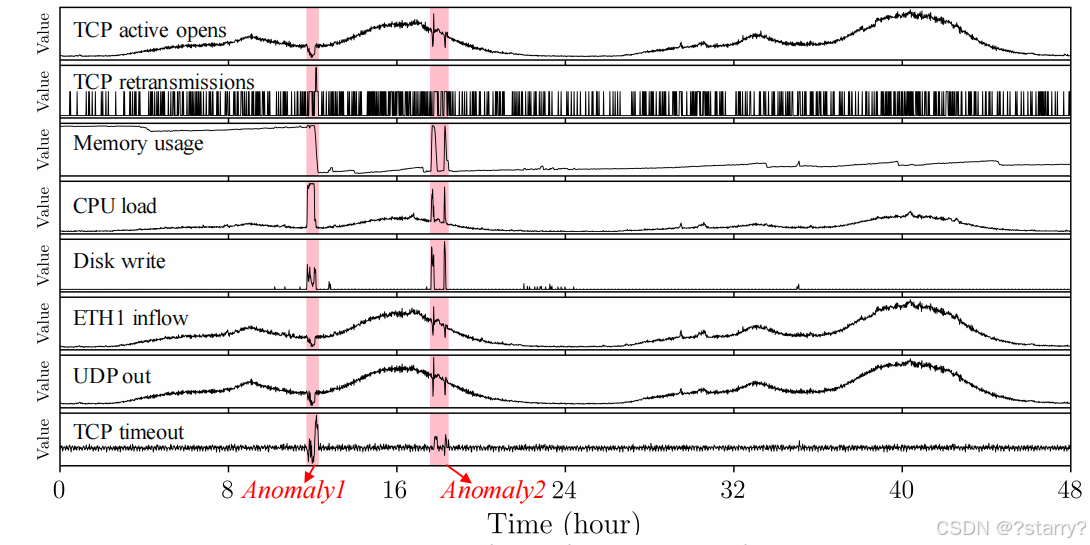
每一个machine-x-y代表一个具体的机器，

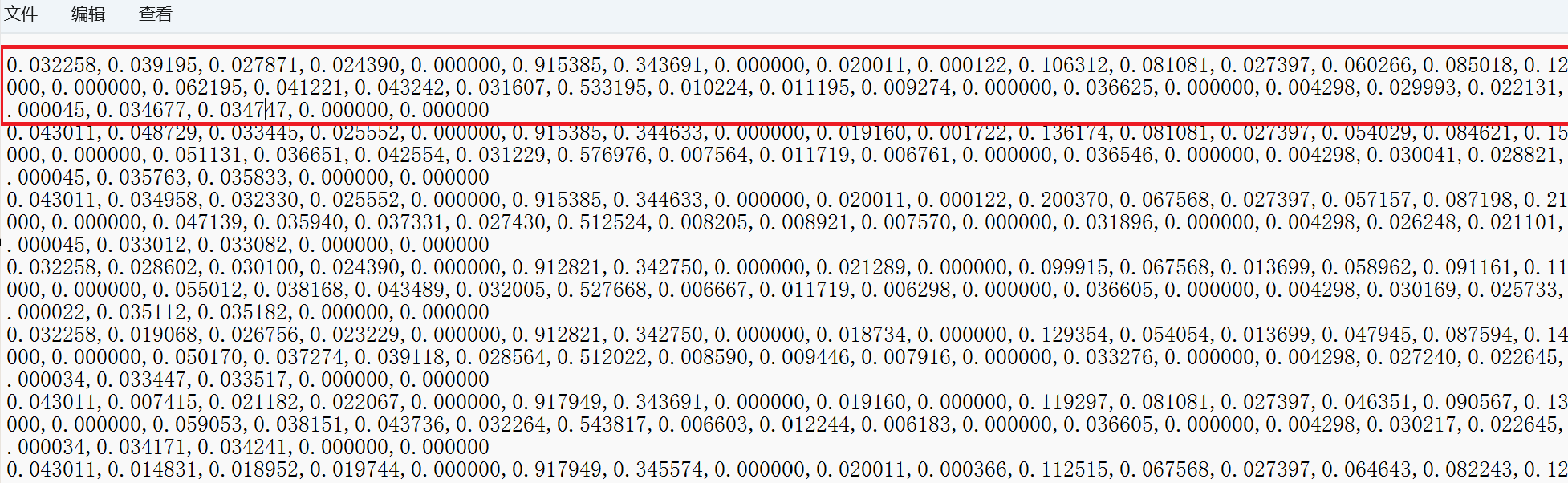
原文件一共四个文件夹。分别是train，test，test\_label，interpretation\_label。

machine-1-1.txt,红圈中即38个维度。

包含了CPU相关指标，内存相关指标，磁盘I/O指标，系统交换指标，TCP相关，UDP协议相关。

其中的八个维度信息如图。

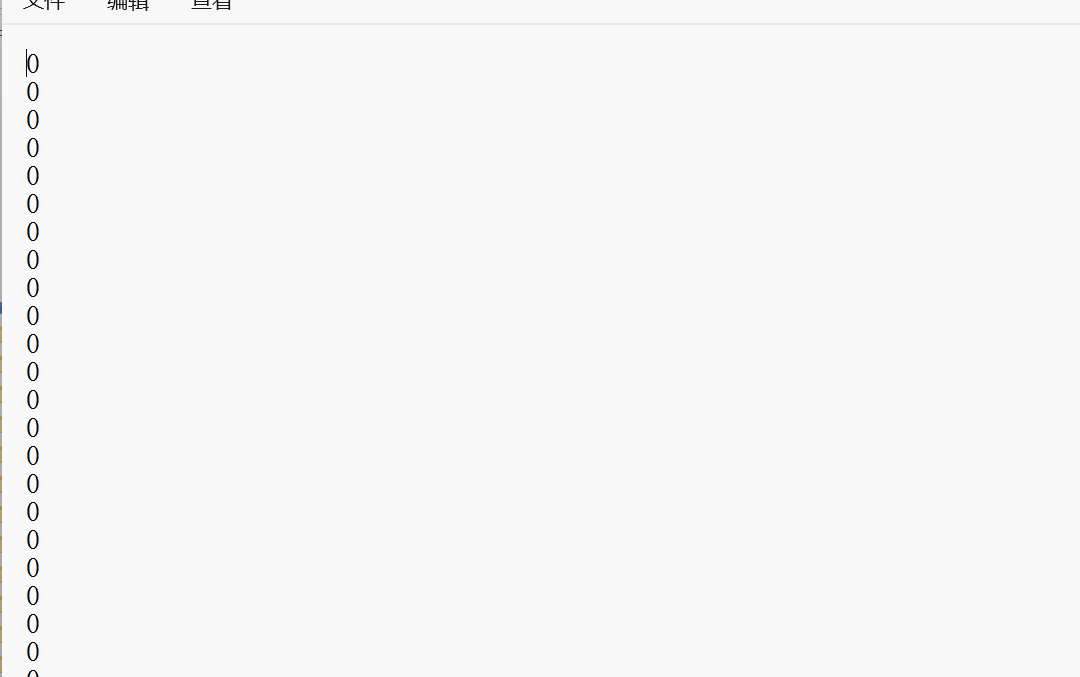




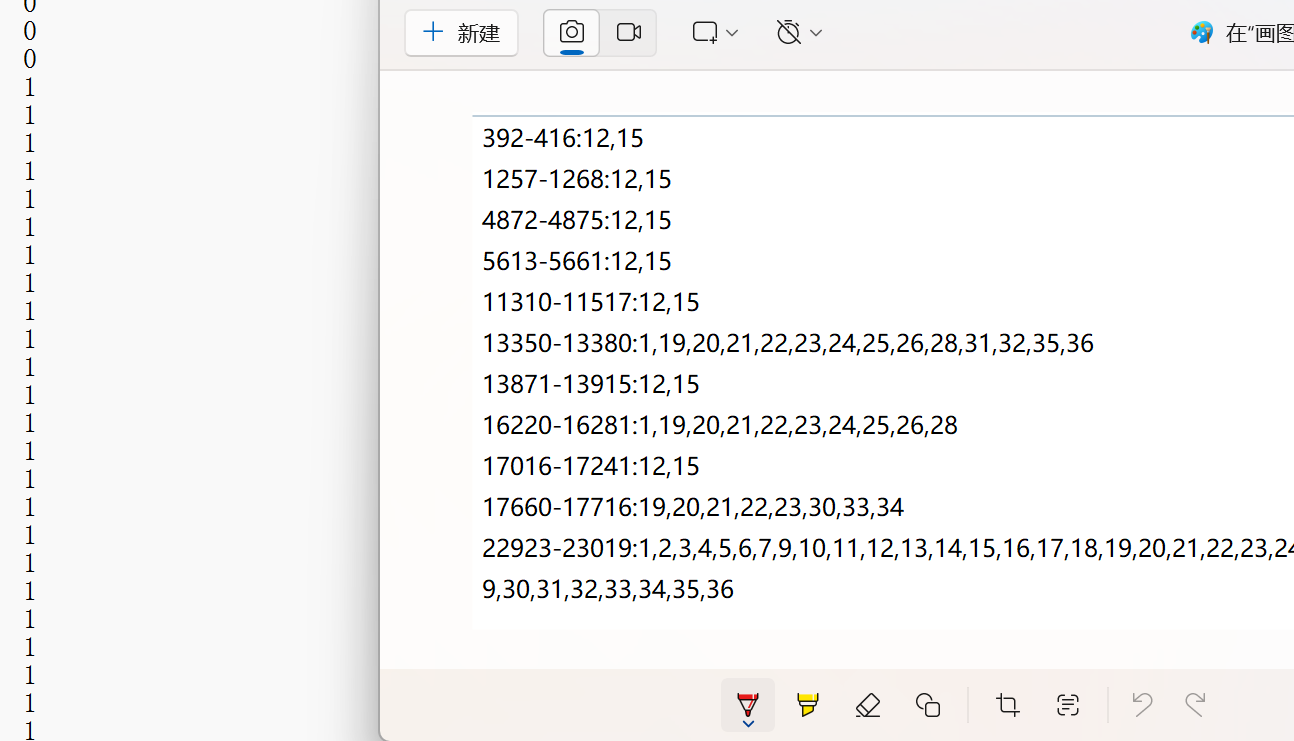
train：训练数据集，无标签。

test：测试数据集，有标签。

test\_label：测试集的便签，0代表正常，1代表异常。



interpretation\_label：对于测试集中的异常数据，哪几个数据特征造成了最终的异常,只要有一个特征异常，本次测量就为异常，已验证interpretation\_label中的异常区间和label中的异常值为1匹配，如下图右边的interpretation\_label 中的392-416中的12,15特征异常，对应左边label中的392-416行值为1。



interpretation\_label

**3.预处理后的文件**

训练集一共708405条数据，测试集一共708420条数据。预处理的文件放在了processed\_csv文件夹下。

处理后的文件命名：machine-x-y\_train.csv ，machine-x-y\_test.csv

X与y代表第x组的第y个机器的特征。

训练和测试数据的列名：

timestamp,col\_0,col\_1,col\_2,col\_3,col\_4,col\_5,col\_6,col\_7,col\_8,col\_9,col\_10,col\_11,col\_12,col\_13,col\_14,col\_15,col\_16,col\_17,col\_18,col\_19,col\_20,col\_21,col\_22,col\_23,col\_24,col\_25,col\_26,col\_27,col\_28,col\_29,col\_30,col\_31,col\_32,col\_33,col\_34,col\_35,col\_36,col\_37

**SMAP & MSL**

* + SMAP: 土壤湿度主动被动卫星(Soil Moisture Active Passive satellite)
  + MSL: 火星科学实验室流浪者(Mars Science Laboratory rover)

SMAP和MSL代表两个不同的产生遥测流的航天器，两个数据集分开

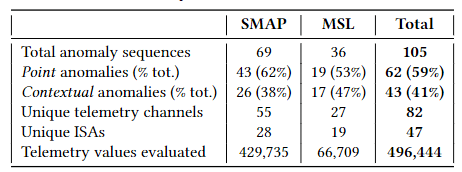
SMAP有55个channel（实体），每个channel有25个维度

MSL有27个channel（实体），每个channel有55个维度

SMAP训练集不含异常数据，训练集的异常数据都由专家打上了异常标签。已验证SMAP包含了562,800个样本，训练集135183，测试集427617。其中异常样本所占比率为 13.13%。

MSL训练集不含异常数据，训练集的异常数据都由专家打上了异常标签。MSL 的样本总共有 132,046 个，训练集58317，测试集73729，异常样本占比为 10.72%。

下表是异常序列以及异常比例的展示。



**2.原文件**

MSL数据集与SMAP数据集在一个文件夹data下。

labeled\_anomalies.csv的列信息：

* chan\_id：匿名通道 ID - 第一个字母代表通道的性质（P = 功率，R = 辐射等），对应于train和test中对应名字的numpy文件。
* Spacecraft：生成遥测流的航天器。MSL or SMAP。
* anomaly\_sequences：流中真实异常的开始和结束索引，对应test中chan\_id文件中的index为此范围的为异常序列。
* Class：异常分类。Point or contextual。
* num\_values：每个流中遥测值的数量，对应chan\_id的test文件的时间戳数量。

**SMAP预处理后的文件**

预处理后的数据集放在了SMAP目录下。为了后续工作能更好的开展，也有将55个channel的数据分开的数据集。

训练和测试数据的列名：

timestamp,col\_0,col\_1,col\_2,col\_3,col\_4,col\_5,col\_6,col\_7,col\_8,col\_9,col\_10,col\_11,col\_12,col\_13,col\_14,col\_15,col\_16,col\_17,col\_18,col\_19,col\_20,col\_21,col\_22,col\_23,col\_

**WADI**

**WADI 数据集 均是 新加坡科技设计大学SUTD的iTrust机构采集并开源的，可在**[**iTrust官网**](https://itrust.sutd.edu.sg/)**申请数据集，申请通过后会向你的邮箱发送Google网盘链接。**

水分配系统数据集(Water Distribution，WaDi)

出处：https://itrust.sutd.edu.sg/itrust-labs\_datasets/dataset\_info/

由Singapore University of Technology and Design收集。来自123个传感器和执行器的数据，一共127个特征。时间间隔1s。

一共运行16天，14天正常运行，2天遭到了15次攻击。异常情景是被攻击的时间收集的传感器数据。

WADI有新和旧两个数据集，分别是WADI.A1\_9 Oct 2017（旧）和WADI.A2\_19 Nov 2019（新）。这是因为由于工厂在运行过程中的某些时段不稳定，需要删除受影响的读数，在2019 年 12 月 19 日删除了受影响的读数，上传了新的数据文件WADI\_14days\_new.csv（在WADI.A2\_19 Nov 2019文件夹下）。

WADI.A2\_19 Nov 2019训练集一共有784570条数据，测试集一共172800条数据。

WADI.A1\_9 Oct 2017训练集一共有1209600条数据，测试集一共172800条数据。

攻击场景：源自研究团队开发的攻击模型。攻击模型将CPS的意图空间(the intent space)视为攻击模型。两天内发起了15次攻击。

**2.原文件**

WADI.A1\_9 Oct 2017这个文件夹下有三个文件，分别是table\_WADI、WADI\_14days.csv和WADI\_attackdata.csv。

table\_WADI：记录了被攻击的时间段。

WADI\_14days.csv：14天正常运行（训练集）。

WADI\_attackdata.csv：2天15次攻击（测试集）。

(784571, 130) (172803, 131) (1209601, 130) (172801, 130)

WADI.A2\_19 Nov 2019这个文件夹下有三个文件，分别是table\_WADI、WADI\_14days\_new.csv和WADI\_attackdataLABLE.csv。

table\_WADI：记录了被攻击的时间段。

WADI\_14days\_new.csv：14天正常运行（训练集），工厂在运行过程中的某些时段不稳定，删除了受影响的读数。

WADI\_attackdataLABLE.csv：2天15次攻击（测试集）。

**3.预处理后的文件**

训练集一共有784570条数据，测试集一共172800条数据。使用的是WADI.A2\_19 Nov 2019（新）标准。

训练集 (784571, 128) 测试集 (172803, 129)

128= 127 特征数据 + 1时间数据

129= 127特征数据 + 1 标签 + 1时间数据