大数据-11.29

我讲的是第四部分-数据获取与比对

这一部分可以继续细分为以下四个部分展开

1是 数据获取的目标，通过论文的参阅找出需要获取到的数据

2是 原始数据集的选择，在论文提供的数据集中选取需要的部分数据

3是 数据的处理 描述具体对于数据集处理从而获取目标数据的流程

最后是 数据的展示比较 通过对获得数据进行可视化用于观察比较

在第一部分的数据获取目标上，参引自论文中对于异常检测部分提到的如下内容：对于给定机器资源的矩阵（也就是选取的特征组合），通过异常检测的方法输出对应的异常分数。如果一个机器的异常分数越小，那么对应于他是异常节点的可能性就越大，而若异常分数小于0，则可以认定对应机器为异常节点。结合论文中给定的参考图，可以得出，在数据获取上，最后需要获取的内容即为每个machine_id与其对应的anomaly score。在获取anomaly score的特征选取上沿用论文中给定的CPU+内存+磁盘的特征组合，结合iForest的方法，对数据进行处理，获取目标数据。

明确数据处理目标与方法以后，在第二部分的原始数据集的选取上，结合论文参考部分提供的github地址，在阿里的仓库下选取了2018年的数据集，具体数据表的各字段在左边这张图上，刚刚提及的特征组合就是cpu_util_percent和mem_util_percent再加一个disk_io_percent三个字段组合。

第三部分是对数据的处理部分，在解压压缩包后获取到一个csv文件，通过观察读取到的数据可以看到原有表格的两个字段存在着数据的缺失，但是实际上这里四个字段都不在选取的特征组合里，在后续就可以一并排除掉了。

处理后剩余的字段就是这样的效果，后续处理上采用两种方案，一种是单台机器集中读取并处理全部80台机器，记录的情况参见左图，通过右边的代码的处理后就可获取到相应数据。另一种方案是分布式地每次单独处理20台机器，最后将4个数据集合合并，一会整体做可视化比对。

最后一部分就是数据的比对部分，在第一组数据，上面蓝色标注的部分是总体80台，黄色的是4次20台的合并，虽然具体每台机器的异常分差异比较大，但是在折线图上可以看出整体曲线上是比较相近的，差别较大的部分通常是在于每20台划分的节点部分，比如1993这个点和1952这一块。

第二组数据里通过之前disk_io_percent字段的异常值判断出异常记录，并找出前10个最能贴合所有异常记录的字段组合，最后得出的异常分状况如左图，可以看出一部分曲线的形态上差异还是比较大的，最能贴合异常记录的字段组合未必能很好地给出所有机器的异常判断结果。

最后将除了两个存在数据缺失的字段排除后，对于剩余5个字段做排列组合后得到31组曲线，可以看到相当部分的曲线其形状是相当接近的，可以通过在这一部分曲线中寻找比对来确定更为优秀的字段组合。

以上就是介绍的全部内容。