异常检测

**项目背景**

随着信息时代硬件计算能力的不断提升，大型公司的客户目标数会急剧增加，而他们的服务器系统必将趋向复杂化，集群化。对于这样的发展前景，服务器系统中的一些微小的事故或异常所引发的“蝴蝶效应”的影响是巨大的，而对于如此大型且复杂的系统，若是凭人力维护，其成本花销也是巨大的。将异常检测算法应用到服务器状态的异常检测能有效降低人力维护成本，但此前的研究成果多限于传统机器学习算法，而对于日趋复杂的系统，这些算法的实用性也逐渐降低（主要在于复杂度的指数级提升），此外样本集的极度不均衡问题（服务器的正常状态远多于异常状态）也导致的基于概率密度的传统方法的实效，因此此项目旨在用生成对抗式网络（GAN）解决样本的不均衡问题并进行高效的异常检测及诊断。

**调研内容**

调研进展目前主要分为三部分，第一部分关于不引入GAN时的较为传统的异常检测方法，第二部分是关于将集成学习与GAN结合的算法及模型，第三部分是关于将半监督学习与GAN结合的模型。（对基本GAN模型不做赘述）

Clustering 半监督：

主要思想为先对训练样本做聚类处理（根据历史KPI曲线形状），并选择每个类的类别中心进行异常标注。新曲线出现后根据其形状将其分配到匹配的类中，若是没有匹配的类则新建一个类。再利用新曲线的数据以及历史上的已有异常标注的类别中心数据训练半监督模型（CPLE），为新KPI曲线做异常检测。

CPLE（Contrastive Pessimistic Likelihood Estimation）是针对有监督模型的算法架构，可以有效利用未标注数据进行更加精确的训练，其流程如下：对已有标注样本训练一个模型来为未标注样本做异常预测，然后用新曲线以及预测的概率值和原曲线一起重新调整模型的优化方向，此过程将会一直迭代重复直到达到指定的迭代次数或未标注曲线的预测值不再发生变化。通过CPLE框架能有效利用未标注样本的特征来指导模型参数的优化，对模型进行自增强，实现监督学习与无监督学习的结合。

集成学习 GAN：

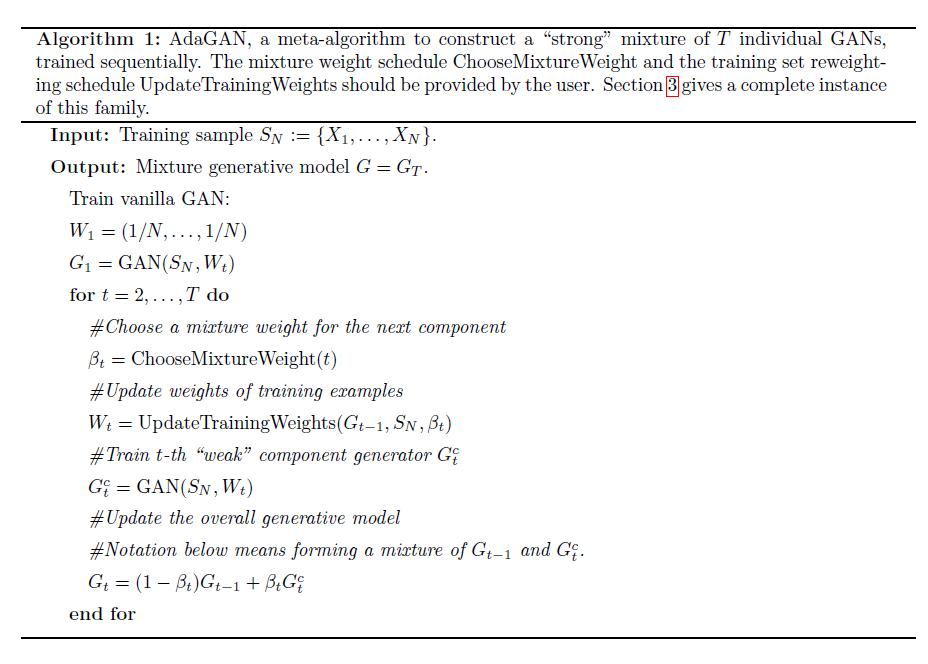
1. Bagging思想：

利用此思想的模型主要分为两类，第一类是多个生成器与单个判别器做对抗训练；第二类是单个生成器和多个判别器做对抗训练。两类的训练机制都较为相似，对于第一类，在每个训练Epoch中随机选取一个生成器的生成结果交由判别器判断，随后将梯度反向传给所有生成器指导它们的后续生成。实验表明此类模型的生成样本质量相对于原始GAN会有所提升。

此外在以上模型的基础上进行改进，a) 设置单个判别器为一个多分类器，对多个生成器分别进行不同的梯度传播，指导生成器学习样本的不同分布，b) 训练多个判别器分别对单个生成器进行梯度回传从而实现对样本不同分布的学习。可以在一定程度上缓解Mode Collapse问题。

2. Boosting思想：

利用Adaboost 的思想，具体做法是在每一个训练epoch结束后训练多个弱生成器，取其中生成效果最好的生成器与现有生成器进行加权，组合成为一个新的强生成器指导下一个epoch的样本生成。实验表明这种模型有利于生成结果的快速收敛，能显著降低运行时耗。



半监督学习 GAN：

对于服务器的异常检测，引入GAN的目的就是为了解决样本不均衡问题—利用GAN的分布拟合能力去拟合正常状态的分布（正常状态样本在样本集中占绝大多数），再通过比较一个测试样本与GAN所学习到的样本分布的差距来判断测试样本是否为异常样本。因为在训练阶段输入样本全为负样本（正常状态样本），可认为此阶段为有监督的学习，而在测试阶段输入样本包含正负样本且无标签（根据样本与学习到的分布差距判断异常与否），可认为此阶段为无监督学习[GANomaly]，因此整个训练及测试过程可看做在GAN框架下的半监督学习过程。

**调研总结**

目前GAN与集成学习结合应用到异常检测的工作基本还没有，同时感觉已有的工作对两个思想的结合较为直接，实验结果相对于原始GAN模型也并不是有很显著的提升（侧重点主要在于模型的收敛速度和Mode Collapse问题的解决）。

而GAN与半监督学习结合的模型[GANomaly]也只是针对图片集的异常检测，对于服务器的异常检测（样本feature存在离散值，不同于图片集feature的连续值）这一类样本集并没有提出相应的优化措施。