# 《Cross-dataset Time Series Anomaly Detection for Cloud Systems》USENIX ATC 2019 微软团队

操作系统、体系结构方面最好的会议

## Abstract

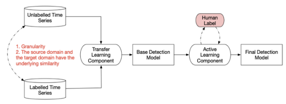
随着云服务的发展，检测云系统的中的异常非常重要，但是由于数据的多样性和巨大量级，很难获得足够的标记数据来建立准确的异常检测模型，因此本文便提出了利用迁移学习和主动学习的ATAD（Active Transfer Anomaly Detection)模型，实验表明效果显著，只需要提供1%-5%的未标记数据进行标记即可。

## Introduciton

- 由于云系统服务的多样性，因此在不同类型的数据之间，异常的定义也是不同的，因此静态阈值是不可取的。
- 基于机器学习的方法，有监督算法需要大量的标签，难以获得，无监督的方法表现不佳
- 因此提出ATAD，充分利用已有标签的其他相似数据集
  - 迁移学习模块
  - 主动学习模块

## Proposed Approach

### 整体框架图




- Unlabelled Time Series：表示尚未标注的时序数据，大都来自内部的监控系统（Target Domain）
- Labelled Time Series：表示开源的已经标注的时序数据（Source Domain）
- Transfer Learning Component：表示将 Source Domain 的特征 迁移到 Target Domain
- Human Label：表示将系统筛选出来的部分样本需要进行人工标注

### Transfer Learning Component

#### 保留时序数据的时间依赖性

**特征提取**

- 在每个数据点提取时间序列相关的特征。将每个数据点从原始的一维标量转换为高维特征向量，并通过这些特征向量保存上下文信息。
- 在该算法中，我们不仅考虑了简单的描述性特征(统计值)，还考虑了顺序感知特征(预测误差特征和时间特征)。

Statistical features

Table 1: Statistical Features

| Feature | Description |
|---|---|
| Mean | Mean. |
| Var | Variance. |
| Crossingpoint[18] | The number of crossing points. |
| ACF1 | First order of autocorrelation. |
| ACFremainder | Autocorrelation of remainder. |
| Trend | Strength of trend. |
| Linearity [18] | Strength of linearity computed on trend of STL [5] decomposition. |
| Curvature [18] | Strength of curvature computed on trend of STL [5] decomposition. |
| Entropy [18] | Spectral entropy [12]. |
| ARCHtest.p [9] | P value of Lagrange Multiplier (LM) test for ARCH model [8]. |
| GARCHtest.p [24] | P value of Lagrange Multiplier (LM) test for GARCH model [8]. |

Forecastingerrorfeatures

Table 2: Metrics used as forecasting error features

| Features | Formula | Description |
|---|---|---|
| $ME$ | $\frac{\sum(Y_i-\hat{Y}_i)}{N}$ | Mean Error. |
| $RMSE$ | $\sqrt{\frac{\sum(Y_i-\hat{Y}_i)^2}{N}}$ | Root Mean Squared Error. |
| $MAE$ | $\frac{\sum\lvert Y_i-\hat{Y}_i\rvert}{N}$ | Mean Absolute Error. |
| $MPE$ | $\frac{1}{N}\cdot\sum\frac{Y_i-\hat{Y}_i}{Y}$ | Mean Percentage Error. |
| $MAPE$ | $\frac{1}{N}\cdot\sum\frac{\lvert Y_i-\hat{Y}_i\rvert}{Y}$ | Mean Average Percentage Error. |

Temporal features

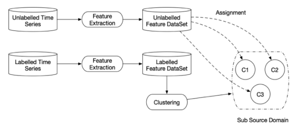
Table 3: Temporal Features

| Features | Description |
|---|---|
| Max_level_shift | Max trimmed mean between two consecutive windows. |
| Max_var_shift | Max variance shift between two consecutive windows. |
| Max_KL_shift | Max shift in Kullback-Leibler divergence between two consecutive windows. |
| Lumpiness | Changing variance in remainder. |
| Flatspots | Discretize time series values into ten equal-sized intervals. Find maximum run length within the same bucket. [18]. |
| Diff-$w$ | The differences between the current value and the $w$-th previous value. |

为了使它们在不同的时间序列之间具有可比性，所有这些特征都被标准化了

#### 粒度级别的一致性

- 时序数据粒度是个需要考虑的因素，比如：Source Domain是天粒度的数据，而 Target Domain是秒或者是分钟粒度的数据，这个是迁移是很难成功的
- 在Source Domain和Target Domain中需要有一定的潜在的相似性，这种相似性是使迁移成为可能的基础




#### 保证源域和目标领域之间数据的相似性

- 聚类方法
  - 将Source域的特征，迁移到Target域中的基本流程，时序数据经过相同的特征工程后，转换成特征，将Source域的特征通过聚类手段（作者使用K-Means算法）分成不同的子域特征，然后将Target域的特征派发到不同的子域中（利用欧式距离进行评估），这样我们就得到了每个子域中对应的特征集合。
- CORrelation ALignment (CORAL)
  - 聚类之后，基于特征的分布可能仍然存在差异，需要从特征空间的角度进一步区分目标域和对应子源域
  - 这里使用的是Correlation Alignment的技术，具体的变换逻辑如下

$$min_A||A^T C_l^i A - C_u^i||_F^2$$

  - 这个公示也很好理解，l表示labelled的特征，u表示unlabelled的特征，i表示某个聚类（子域），A是我们要寻找的一个线性变换矩阵。通过上述公式，我们可以找到一个A矩阵，我们通过A矩阵，可以将Source域的特征，转换到Target域中。

在拿到相同域的特征后，可以使用简单的分类模型对尚未明确标签的特征进行打标签，作者使用的RF模型作为基础的分类器。

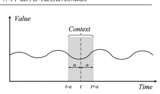
### Active Learning Component

主动学习的目的是要将基础的分类器打分的样本进行筛选，将筛选出来的样本按照一定的规则分发给标注人员，进行人工打标的操作，那么如何筛选就是一个需要解决的问题

最常用的不确定性准则 $Uncertainty = -|Prob(Normal) - Prob(Anomaly)|$

多样性冗余

- 推荐标记样品时，多样性也是需要考虑的一个重要因素。有时，两个样本非常相似或可能属于相同的异常模式。没有必要同时给它们贴上标签。
- 具体来说，我们根据不确定性对所有样本进行排序，并按顺序扫描它们。如果我们扫描的一个新样本在候选集中的另一个样本的文本中，即这两个样本是相邻的，我们认为它们所嵌入的信息也可以是相似的。因此，我们忽略了新样本，因为它可能包含冗余信息。




通过这上述步骤后，再去训练基础模型，如此重复2～3次后可以调优模型，达到最终的Final Model

## Experiment

在本节中，我们通过一系列的实验来评估我们的方法的有效性。我们的目的是在评估中回答以下研究问题:

### RQ1:建议的ATAD方法有多有效?

Table 5: Results of Comparative Methods

| Dataset | Method | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Non-Yahoo → Yahoo | iForest | 0.3832 | 0.2183 | 0.2781 |
| | K-Sigma | 0.6499 | 0.3364 | 0.4433 |
| | S-H-ESD | 0.2779 | 0.6215 | 0.3840 |
| | RF | 0.8668 | 0.2075 | 0.3348 |
| | ATAD | 0.8847 | 0.4040 | **0.5547** |
| Non-AWS → AWS | iForest | 0.1523 | 0.0491 | 0.0743 |
| | K-Sigma | 0.6899 | 0.1992 | 0.3091 |
| | S-H-ESD | 0.5382 | 0.7100 | 0.6123 |
| | RF | 0.9999 | 0.6226 | 0.7674 |
| | ATAD | 0.9195 | 0.8142 | **0.8637** |
| Non-Artificial → Artificial | iForest | 0.3477 | 0.9006 | 0.5017 |
| | K-Sigma | 1.000 | 0.1730 | 0.2950 |
| | S-H-ESD | 0.7888 | 0.4568 | 0.5785 |
| | RF | 0.9182 | 0.9301 | 0.9241 |
| | ATAD | 0.9990 | 0.9850 | **0.9924** |
| Non-Twitter → Twitter | iForest | 0.4685 | 0.3087 | 0.3722 |
| | K-Sigma | 0.2608 | 1.0000 | 0.2608 |
| | S-H-ESD | 0.7481 | 0.4654 | 0.5739 |
| | RF | 0.7285 | 0.4811 | 0.5795 |
| | ATAD | 0.8769 | 0.6951 | **0.7755** |

Table 7: The effectiveness of features (F1-Score)

| Features | Yahoo | AWS | Artifical | Twitter |
|---|---|---|---|---|
| Statistical | 0.2956 | 0.7387 | 0.7441 | 0.6937 |
| Order-aware | 0.4200 | 0.8441 | 0.7569 | 0.6622 |
| All features | **0.5781** | **0.8637** | **0.9924** | **0.7755** |

### RQ2:迁移学习部分的效果如何?

Table 8: Feature Importance Evaluation

| Dataset | Important Features |
|---|---|
| Yahoo | Original_data, RMSE, MAE, ME, Mean, MPE, Diff-$p$, Diff-$w_{pr/2}$, Diff-2$p$, MAPE |
| AWS | Original_data, RMSE, MAE, ME, Mean, ACF1, Diff-2$p$, Curvature, Flatspots, Diff-$p$ |
| Artificial | Original_data, Mean, Diff-$p$, MPE, Flatspots, RMSE, MAE, Diff-2$p$, Diff-$w_{pr}$, Lumpiness. |
| Twitter | Var, RMSE, MAE, ARCHtest.p, Mean, Flatspots, Original_data, Max_level_shift, MPE, Original Data |

Table 9: Comparative Experiment of ATAD and Naive Active Learning without Transfer Learning (F1-Score)

| | Naive F1-Score | Naive #labels | ATAD F1-Score | ATAD #labels |
|---|---|---|---|---|
| Yahoo | 0.5691 | 1380 | 0.5697 | 920 |
| AWS | 0.8589 | 381 | 0.8637 | 254 |
| Artificial | 0.9815 | 322 | 0.9924 | 161 |
| Twitter | 0.6164 | 2855 | 0.7755 | 1427 |

### RQ3:主动学习的效果如何?

Table 10: The experimental result with active learning alone (F1-Score)

| Dataset | Method | Base | Round1 | Round2 | Round3 |
|---|---|---|---|---|---|
| Yahoo | U | 0.2090 | 0.2437 | 0.2839 | 0.2695 |
| | UCD | 0.2090 | 0.2383 | 0.3032 | **0.3814** |
| | random | 0.2090 | 0.2275 | 0.2256 | 0.2196 |
| AWS | U | 0.0443 | 0.1932 | 0.6838 | 0.6799 |
| | UCD | 0.0443 | 0.7550 | 0.8092 | **0.8879** |
| | random | 0.0443 | 0.3227 | 0.5164 | 0.6604 |
| Artificial | U | 0.6239 | 0.7272 | 0.8737 | 0.9006 |
| | UCD | 0.6239 | 0.7340 | 0.8780 | **0.9715** |
| | random | 0.6239 | 0.6446 | 0.6275 | 0.5000 |
| Twitter | U | 0.1647 | 0.2916 | 0.2959 | 0.3298 |
| | UCD | 0.1999 | 0.3070 | 0.3703 | **0.4232** |
| | random | 0.1647 | 0.1798 | 0.1944 | 0.1689 |

### RQ4: ATAD检测异常的有效性?基于公共数据集的公司的本地数据集?

Table 11: Experimental result on IOPS dataset of Microsoft

| | Precision | Recall | F1-Score |
|---|---|---|---|
| iForest | 0.2886 | 0.3988 | 0.3349 |
| K-Sigma | 0.8170 | 0.1882 | 0.3059 |
| S-H-ESD | 0.9117 | 0.1741 | 0.2924 |
| RF | 0.5213 | 0.6724 | 0.5873 |
| ATAD | 0.8082 | 0.6188 | **0.7009** |

### 数据集

Table 4: Summary of datasets

| Dataset | time series mean length | #data points | #anomaly points | %anomaly |
|---|---|---|---|---|
| Yahoo | 1415 | 92016 | 1617 | 1.76% |
| AWS | 3985 | 67740 | 3097 | 4.57% |
| Artificial | 4032 | 16128 | 624 | 3.87% |
| Twitter | 15862 | 142765 | 217 | 0.15% |