大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

トップエスイー ソフトウェア開発実践演習

ログ分析における マイクロサービスシステムの 異常検知技術に関する演習

佐々木拓哉(株式会社NTTデータ)　加藤雅也(富士通株式会社)
定行裕輔(NTTテクノクロス株式会社)　内田柊平(株式会社NTTデータアイ)

システムログの異常検知技術における問題点

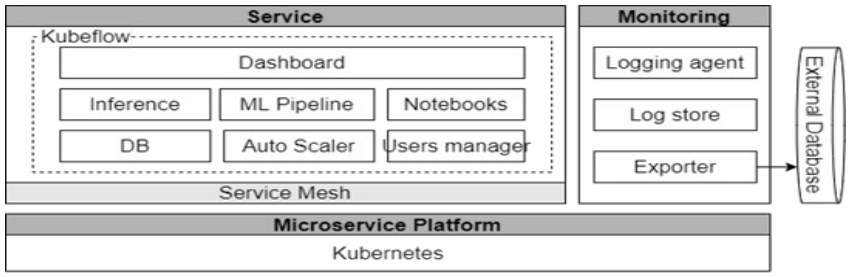
モデルの評価用データセットの多様性の不足
1. 複雑な構造が想定されるマイクロサービスの オープンソースのデータセットが存在しない
2. ログのカテゴリに応じた改善方針の議論が 少なくマイクロサービス固有の改善が不明

手法・ツールの適用による解決

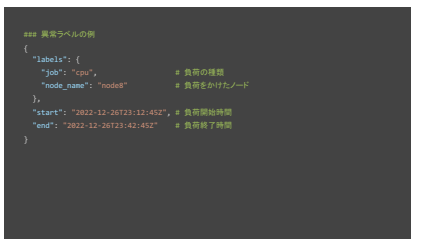
1. システムの構築とデータセットの作成
   ・PLGスタックを含んだKubeflowの構築
   ・異常データの発生シナリオの検討
2. 作成したデータセットの評価結果に基づいた 分析プロセスの改善を実施
   ・グルーピングと呼ばれる処理に着目

データセット

●システム構築
マイクロサービスのログを収集するためにシステムを構築した。



●負荷をかけてログを収集
①CPU使用率が95%以上
②メモリ使用率が95%以上
③ノードダウン
④ディスク使用率が100%



Logbertを使用した学習

●テンプレート作成
ログメッセージ内に含まれるタイムスタンプやログレベルなどのパラメータをマスクしテンプレートを作成した。

2022-12-24T10:48:24 info ads RDS: PUSH for node:metadata-writer-8bd8b7b66-chjk7.kubeflow resources:34 size:49.5kB

<*> info ads RDS: PUSH for node<*>metadata-writer<*> resources:<*> size:<*>

●グルーピング
ノードIDを識別子としてログデータをノードごとにグループ化し、固定長のログシーケンスを作成しました。

時刻1：ログメッセージ(Node1)
時刻3：ログメッセージ(Node1)
時刻6：ログメッセージ(Node1)

時刻2：ログメッセージ(Node2)
時刻5：ログメッセージ(Node2)

時刻4：ログメッセージ(Node3)

異常検知結果

シーケンス数	適合率	再現率	F値
グルーピングなし セッションサイズ:40	0.857	0.551	0.671
グルーピングなし セッションサイズ:100	0.827	0.728	0.775
グルーピングあり セッションサイズ:40	0.896	0.762	0.823
グルーピングあり セッションサイズ:100	0.870	0.810	0.840

ノード単位のグループ分けを行うことで精度の向上がみられ、F値0.80程度のスコアを出すことができた。

今後の課題

● ハイパーパラメータの調整
今回の演習ではハイパーパラメータに関する調査が不十分でした。各種パラメータを調整することで精度改善につながると考えている。

● 今回深く言及できなかったデータセット
作成したすべてのデータセットを試すことができなかったため、今後はそのデータに対する分析が必要だと考えている。

● グルーピングの改善
作成したログには、グルーピングに使用できる情報ノード以外も含まれているため、それらも活用できると考えている。

● スペシャルトークンの活用
特定のログキーの出力状況を反映したスペシャルトークンを導入することで、精度改善の可能性があると考えている。