# インテル® DAAL を使用した外れ値検出の強化

この記事は、インテル® デベロッパー・ゾーンに公開されている「Enhancing Outlier Detection with Intel® DAAL」の日本語参考訳です。

---

## はじめに

クレジットカード会社は、どのように不正や悪用を検出しているのでしょうか? ネットワーク管理者は、どのように侵入を発見しているのでしょうか? 科学者は、実験が正しく行われたかどうかをどのように知ることができるのでしょうか?

これらを行うためには、データセットを分析して正規性から外れたデータポイントを探します。例えば、クレジットカード会社は、特定の取引での異常な高額請求や、奇妙な購買行動を探します。これらは、クレジットカードが盗まれたことを示している可能性があります。ネットワーク管理者は、ネットワーク侵入の可能性を示す、特定の場所からの異常な負荷や国外の IP アドレスからのネットワーク・アクセスなどの不規則なアクティビティーをログファイルで探します。同様に、科学者は、実験が正しく行われていないことを示す指標として、正常な範囲または想定範囲から外れたデータを探します。

このような異常なアクティビティーや不規則なアクティビティーは、外れ値または異常値と呼ばれます。この記事では、データ内の外れ値 [1] を検出するさまざまな方法について説明し、インテル® データ・アナリティクス・アクセラレーション・ライブラリー (インテル® DAAL)[2] を使用してインテル® Xeon® プロセッサー・ベースのシステム向けに外れ値の検出を最適化する方法を紹介します。

## 外れ値とは?

外れ値は、ほかのデータから大きく異なる (異常または不規則)、または乖離しているデータポイントです (図 1 を参照)。

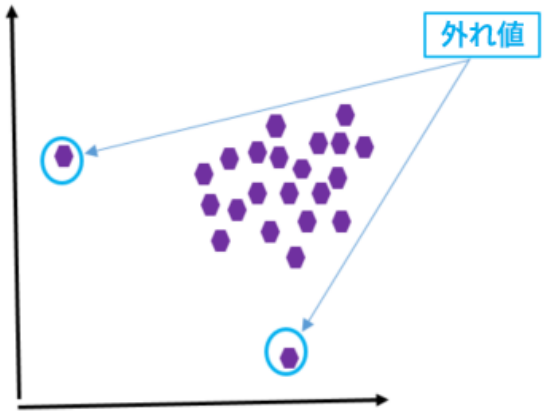


図 1: 外れ値 - ケース #1

紫の点は、データセット内のデータポイントを示します。このグラフでは、ほかのデータポイントから大きく離れている 2 つのデータポイントが外れ値と見なされます。

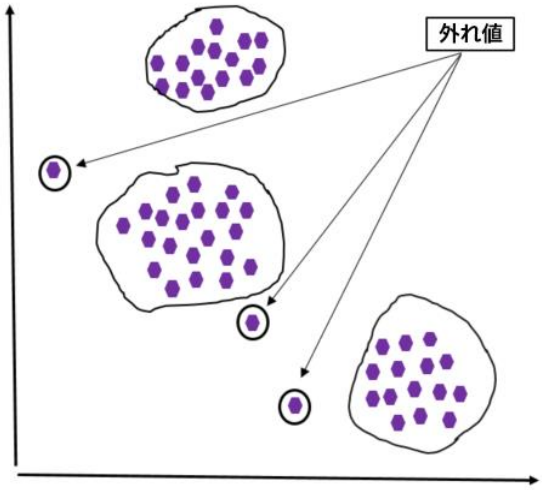


図 2: 外れ値 - ケース #2

図 2 は別の外れ値の例です。このケースでは、データセットは 3 つのグループ (クラスター) に分けることができます。グループの外側のデータポイントはすべて外れ値と見なされます。

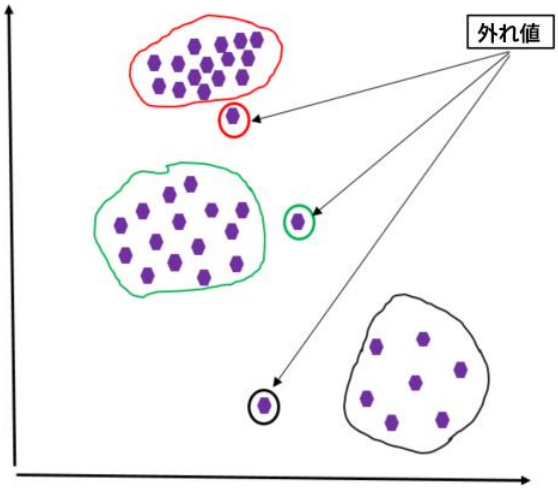


図 3: 外れ値 - ケース #3

図 3 は別の外れ値の例です。ここでも、データセットは 3 つのグループに分けることができますが、このケースは図 2 とは異なります。図2では、すべてのグループでデータポイントがほぼ一様に分布しているのに対し、図3ではグループごとにデータポイントの密度が異なります。

## 外れ値の原因は何か?

外れ値には、メリットとデメリットがあります。不規則なアクティビティー (外れ値) を検出することで、ネットワーク管理者は潜在的な侵入を検出して防ぐことができます。一方で、外れ値を検出して除外することで、計算結果への影響を排除したり、最小限に抑えることができます。外れ値は、マシンラーニング [3] アルゴリズムのトレーニング・プロセスを歪曲したり、誤った方向に導く可能性があり、その結果、トレーニング時間が長くなったり、モデルの精度が低下することがあります。例えば、K 平均法クラスタリング・アルゴリズムでは、データセット内の外れ値は、クラスターのセントロイドを本来の位置から遠ざけます。

一般的な外れ値の原因には、以下のものがあります。

- データ収集エラー: データ収集デバイスは、ノイズが原因で異常なデータを収集してしまうことがあります。
- データ入力エラー: 入力されたデータが正しくありません。例えば、住宅の販売価格を誤入力すると、その住宅の価格はその地域の平均的な住宅価格の範囲外になります。
- 選択エラー: 例えば、高校生の身長について考えてみます。バスケットボール部に所属する一部の生徒の身長は、ほかの生徒と比べて非常に高いため、外れ値になります。バスケットボール部の生徒の身長は、生徒全体とは別に測定すべきです。
- 変換エラー: 複数のソースからデータを抽出する場合、操作ミスや抽出ミスが外れ値の原因になることがあります。

## 外れ値の検出方法

外れ値を検出する一般的な方法は、図 1 - 3 のようなデータセットをプロットしたグラフを見ることです。

Charu C. Aggarwal 著『Outlier Analysis』(英語) 第 2 版 [4] では、次の外れ値検出方法が紹介されています。

- 確率モデル
- 線形モデル
- 近似ベースのモデル
- 高次元の外れ値検出

## 外れ値検出の適用例

外れ値検出方法は、奇妙なデータや異常データを検出できるため、次の用途があります。

- ネットワーク・セキュリティー分析で不規則なアクティビティーや奇妙なアドレスを検出
- 異常な購入パターンや非常に高額な取引を監視して、クレジットカードの不正使用を識別
- 異常な症状や検査結果を発見して、患者の潜在的な健康問題を診断
- ほかの選手と比較して異常なデータを分析して、優れたスポーツ選手を識別

これはほんの一部です。外れ値検出方法は、ほかの多くのものに適用できます。

## インテル® DAAL

インテル® DAAL は、データ解析とマシンラーニング向けに最適化された多くの基本ビルディング・ブロックからなるライブラリーです。これらの基本ビルディング・ブロックは、最新のインテル® プロセッサーの機能向けに高度に最適化されています。この記事では、インテル® DAAL の Python* API を使用して、外れ値検出関数を呼び出す方法を示します。インテル® DAAL をインストールするには、ドキュメント [5] の手順に従ってください。

## インテル® DAAL の外れ値検出アルゴリズムを使用する

以下は、インテル® DAAL ドキュメントから抜粋した、単変量外れ値の説明と外れ値領域を定義する式です。

「次元 p の n 特徴ベクトル $x_1 = (x_{11},\cdots,x_{1p})$, ..., $x_n = (x_{n1},\cdots,x_{np})$ のセット X で、分布に属さないベクトルを識別します。単変量外れ値検出のアルゴリズムは、それぞれの特徴を個別に考慮します。単変量外れ値検出法はパラメトリックで、データセットの既知の根本的な分布配置を仮定し、観測点が領域に属している場合に外れ値としてマークする外れ値領域を定義します。外れ値領域の定義は、仮定された根本的なデータ分布に結合されます。単変量外れ値検出の外れ値領域の例を次に示します。

$$Outlier(\alpha_n, m_n, \sigma_n) = \left\{ x : \frac{|x - m_n|}{\sigma_n} > g(n, \alpha_n) \right\},$$

ここで、$m_n$ および $\sigma_n$ は指定されたデータセットについて計算された平均および標準偏差の (ロバスト) 推定、$\alpha_n$ は信頼度係数です。g (n, $\alpha_n$) は領域の範囲を定義し、観測点の数に調整します。」

このセクションでは、インテル® DAAL の Python*[6] 外れ値アルゴリズムを呼び出す方法を示します。

次のステップに従って、インテル® DAAL から単変量外れ値検出アルゴリズムを呼び出します。

1. from コマンドと import コマンドを使用して、必要なパッケージをインポートします。
    1. 次のコマンドを実行して、インテル® DAAL の数値テーブルをインポートします。

```
from daal.data_management import FileDataSource, writeOnly,
DataSourceIface, BlockDescriptor_Float64
```

    2. 次のコマンドを実行して、単変量外れ値検出アルゴリズムをインポートします。

```
from daal.algorithms.univariate_outlier_detection import
InitIface, Batch_Float64DefaultDense, data, weights
```

2. データ入力が .csv ファイルからの場合、ファイル・データ・ソースを初期化します。

```
DataSet = FileDataSource(
    trainDatasetFileName, DataSourceIface.doAllocateNumericTable,
    DataSourceIface.doDictionaryFromContext
  )
```

3. 入力データをロードします。

```
DataSet.loadDataBlock()
nFeatures = DataSet.getNumberOfColumns()
```

4. 関数を作成します。
    1. 最初に、アルゴリズム・オブジェクトを作成します。

```
algorithm = Batch_Float64DefaultDense()
```

    2. アルゴリズムにデータセットを渡します。

```
algorithm.input.set(data, DataSet.getNumericTable())
```

5. 外れ値を計算して結果を取得します。

```
results = algorithm.compute()
```

6. 次のコマンドを実行して、結果を出力します。

```
printNumericTable(results.get(weights), "outlier results")
```

注: カリフォルニア大学アーバイン校のマシンラーニング・リポジトリー[7] から一般的なデータセットを利用できます。

## まとめ

外れ値検出は、不正検出、ネットワーク・セキュリティーなどにおいて重要な役割を果たします。インテル® DAAL の外れ値検出アルゴリズムは最適化されています。インテル® DAAL を使用することで、アプリケーションを変更せずに、インテル® DAAL の最新バージョンにリンクするだけで、将来の世代のインテル® Xeon® プロセッサーの新機能を利用できます。

## 関連情報

1. 異常検知
2. インテル® DAAL の概要 (英語)
3. Wikipedia - 機械学習
4. 『Outlier detection』(英語)
5. Linux* でインテル® DAAL の Python* バージョンをインストールする方法 (英語)
6. Python* ウェブサイト (英語)
7. 一般的なデータセット (英語)

---