Anomaly Detection in High Dimensional Data

발표자

정희철

INDEX

1

논문 소개

Anomaly Detection in High Dimensional Data

기존 고차원 이상치탐지 알고리즘인 HDoutliers의 단점을

개선시킨 Stray Algorithm 소개 및 패키지 배포

Anomaly Detection in High Dimensional Data

Priyanga Dilini Talagala[1,3,4]

and

Rob J. Hyndman[1,3]

and

Kate Smith-Miles[2,3]

[1]Department of Econometrics and Business Statistics, Monash University, Australia

[2]School of Mathematics and Statistics, University of Melbourne, Australia

[3]ARC Centre of Excellence for Mathematics and Statistical Frontiers (ACEMS), Australia

[4]Department of Computational Mathematics, University of Moratuwa, Sri Lanka

Corresponding author Priyanga Dilini Talagala priyangad@uom.lk

Abstract

The HDoutliers algorithm is a powerful unsupervised algorithm for detecting anomalies in high-dimensional data, with a strong theoretical foundation. However, it suffers from some limitations that significantly hinder its performance level, under certain circumstances. In this article, we propose an algorithm that addresses these limitations. We define an anomaly as an observation where its k-nearest neighbour distance with the maximum gap is significantly different from what we would expect if the distribution of k-nearest neighbours with the maximum gap is in the maximum domain of attraction of the Gumbel distribution. An approach based on extreme value theory is used for the anomalous threshold calculation. Using various synthetic and real datasets, we demonstrate the wide applicability and usefulness of our algorithm, which we call the stray algorithm. We also demonstrate how this algorithm can assist in detecting anomalies present in other data structures using feature engineering. We show the situations where the stray algorithm outperforms the HDoutliers algorithm both in accuracy and computational time. This framework is implemented in the open source R package `stray`.

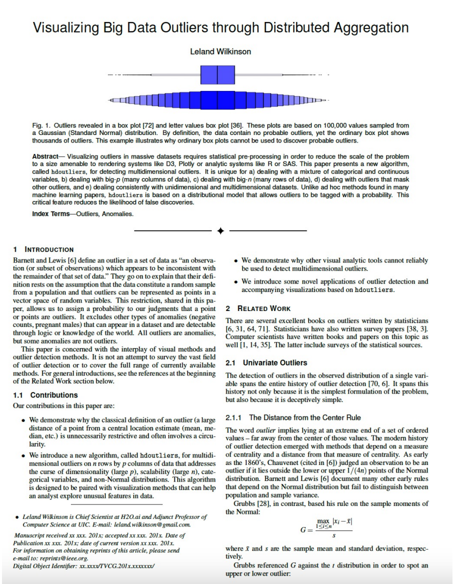
- Priyanga Dilini Talagala
- Department of Econometrics and Business Statistics, Monash University (Australia)
- School of Mathematics and Statistics, University of Melbourne (Australia)
- Department of Computational Mathematics, University of Moratuwa (Sri Lanka)

2

HDoutliers Algorithm

"Visualizing Big Data Outliers through Distributed Aggregation"

거리기반 이상치탐지 기법으로 차원에 상관없이
1차원 문제로 축소시켜 직관적



Visualizing Big Data Outliers through Distributed Aggregation

Leland Wilkinson

Fig. 1. Outliers revealed in a box plot [72] and letter values box plot [36]. These plots are based on 100,000 values sampled from a Gaussian (Standard Normal) distribution. By definition, the data contain no probable outliers, yet the ordinary box plot shows thousands of outliers. This example illustrates why ordinary box plots cannot be used to discover probable outliers.

Abstract— Visualizing outliers in massive datasets requires statistical pre-processing in order to reduce the scale of the problem to a size amenable to rendering systems like D3, Plotly or analytic systems like R or SAS. This paper presents a new algorithm, called `hdoutliers`, for detecting multidimensional outliers. It is unique for a) dealing with a mixture of categorical and continuous variables, b) dealing with big-*p* (many columns of data), c) dealing with big-*n* (many rows of data), d) dealing with outliers that mask other outliers, and e) dealing consistently with unidimensional and multidimensional datasets. Unlike ad hoc methods found in many machine learning papers, `hdoutliers` is based on a distributional model that allows outliers to be tagged with a probability. This critical feature reduces the likelihood of false discoveries.

Index Terms—Outliers, Anomalies.

1 INTRODUCTION

Barnett and Lewis [6] define an outlier in a set of data as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data." They go on to explain that their definition rests on the assumption that the data constitute a random sample from a population and that outliers can be represented as points in a vector space of random variables. This restriction, shared in this paper, allows us to assign a probability to our judgments that a point or points are outliers. It excludes other types of anomalies (negative counts, pregnant males) that can appear in a dataset and are detectable through logic or knowledge of the world. All outliers are anomalies, but some anomalies are not outliers.

This paper is concerned with the interplay of visual methods and outlier detection methods. It is not an attempt to survey the vast field of outlier detection or to cover the full range of currently available methods. For general introductions, see the references at the beginning of the Related Work section below.

1.1 Contributions

Our contributions in this paper are:

- We demonstrate why the classical definition of an outlier (a large distance of a point from a central location estimate (mean, median, etc.) is unnecessarily restrictive and often involves a circularity.
- We introduce a new algorithm, called `hdoutliers`, for multidimensional outliers on n rows by p columns of data that addresses the curse of dimensionality (large p), scalability (large n), categorical variables, and non-Normal distributions. This algorithm is designed to be paired with visualization methods that can help an analyst explore unusual features in data.
- We demonstrate why other visual analytic tools cannot reliably be used to detect multidimensional outliers.
- We introduce some novel applications of outlier detection and accompanying visualizations based on `hdoutliers`.

2 RELATED WORK

There are several excellent books on outliers written by statisticians [6, 31, 64, 71]. Statisticians have also written survey papers [38, 3]. Computer scientists have written books and papers on this topic as well [1, 14, 35]. The latter include surveys of the statistical sources.

2.1 Univariate Outliers

The detection of outliers in the observed distribution of a single variable spans the entire history of outlier detection [70, 6]. It spans this history not only because it is the simplest formulation of the problem, but also because it is deceptively simple.

2.1.1 The Distance from the Center Rule

The word *outlier* implies lying at an extreme end of a set of ordered values – far away from the center of those values. The modern history of outlier detection emerged with methods that depend on a measure of centrality and a distance from that measure of centrality. As early as the 1860's, Chauvenet (cited in [6]) judged an observation to be an outlier if it lies outside the lower or upper $1/(4n)$ points of the Normal distribution. Barnett and Lewis [6] document many other early rules that depend on the Normal distribution but fail to distinguish between population and sample variance.

Grubbs [28], in contrast, based his rule on the sample moments of the Normal:

$$G = \frac{\max_{1 \le i \le n} |x_i - \bar{x}|}{s}$$

where $\bar{x}$ and s are the sample mean and standard deviation, respectively.

Grubbs referenced G against the t distribution in order to spot an upper or lower outlier:

- *Leland Wilkinson is Chief Scientist at H2O.ai and Adjunct Professor of Computer Science at UIC. E-mail: leland.wilkinson@gmail.com.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx/

- Leland Wilkinson (2017)
- Department of Computer Science, University of Illinois , Chicago
- Version 1 : 각 점마다 nearest neighbor distance를 계산하여 이상치탐지
- Version 2 : 선제적으로 클러스터링을 진행하여 각 클러스터마다 대표점을 산출하고, 대표점들만을 사용하여 nearest neighbor distance로 이상치탐지

2 HDoutliers Algorithm

"Visualizing Big Data Outliers through Distributed Aggregation"

거리기반 이상치탐지 기법으로 차원에 상관없이
1차원 문제로 축소시켜 직관적

Visualizing Big Data Outliers through Distributed Aggregation
Leland Wilkinson

- Leland Wilkinson (2017)
- Department of Computer Science, University of Illinois , Chicago
- Version 1 : 각 점마다 nearest neighbor distance를 계산하여 이상치탐지
- ✓ Version 2 : 선제적으로 클러스터링을 진행하여 각 클러스터마다 대표점을 산출하고, 대표점들만을 사용하여 nearest neighbor distance로 이상치탐지

2 HDoutliers Algorithm

HDoutliers Algorithm의 문제점

1

오직 nearest neighbor distance만을 사용하여 이상치탐지

2

대표점 산출을 위한 중간 클러스터링 과정 추가

3

Anomalous Threshold 계산 과정에서
FN rate을 증가시키는 경향이 있음

2 HDoutliers Algorithm

HDoutliers Algorithm의 문제점

1

오직 nearest neighbor distance만을 사용하여 이상치탐지

2

대표점 산출을 위한 중간 클러스터링 과정 추가

3

Anomalous Threshold 계산 과정에서
FN rate을 증가시키는 경향이 있음

HDoutliers Algorithm의 문제점

1

오직 nearest neighbor distance만을 사용하여 이상치탐지

2

대표점 산출을 위한 중간 클러스터링 과정 추가

3

Anomalous Threshold 계산 과정에서

FN rate을 증가시키는 경향이 있음

HDoutliers Algorithm의 문제점

1

오직 nearest neighbor distance만을 사용하여 이상치탐지

필요한 가정 : Isolated Anomalies

- "정상치와 이상치 간의 거리는 멀다" 가정이 충족되어야 함

- 해당 가정 하에, 이상치는 nearest neighbor distance가 매우 큼 (threshold에 관해서는 추후 설명)

- 특정 Representative Point가 nearest neighbor distance가 threshold 보다 크다면 anomalous point로 간주하고, 해당 Representative Point 가 포함된 클러스터 전체가 anomalous points로 판별

HDoutliers Algorithm의 문제점

1

오직 nearest neighbor distance만을 사용하여 이상치탐지

필요한 가정 : Isolated Anomalies

- “정상치와 이상치 간의 거리는 멀다” 가정이 충족되어야 함

- 해당 가정 하에, 이상치는 nearest neighbor distance가 매우 큼 (threshold에 관해서는 추후 설명)

- 특정 Representative Point의 nearest neighbor distance가 threshold 보다 크다면 이상치로 간주하고, 해당 Representative Point가 포함된 클러스터 전체가 이상치로 판별

HDoutliers Algorithm의 문제점

1

오직 nearest neighbor distance만을 사용하여 이상치탐지

파생되는 문제점

2개 이상의 이상치 클러스터들이 서로 가깝게 형성 되어 있다면, 이들 간의 nearest neighbor distance가 작게 산출되면서 정상 클러스터로 판별

- 해당 가정 하에, 이상치는 nearest neighbor distance가 매우 큼 (threshold에 관해서는 추후 설명)

- 특정 Representative Point가 nearest neighbor distance가 threshold 보다 크다면 anomalous point로 간주하고, 해당 Representative Point가 포함된 클러스터 전체가 anomalous points로 판별

HDoutliers Algorithm의 문제점

파생되는 문제점

1

오직 nearest neighbor distance만을 사용하여 이상치탐지

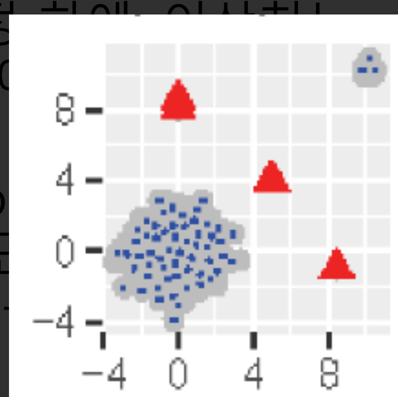
2개 이상의 이상치 클러스터들이 서로 가깝게 형성되어 있다면, 이들 간의 nearest neighbor distance가 작게 산출되면서 정상 클러스터로판별

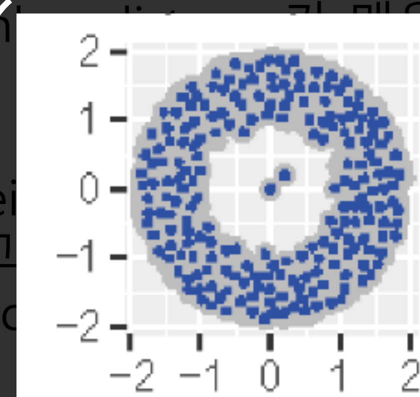
Isolated Anomalies

- 해당 가정 하에, 이상치는 nearest neighbor distance가 매우 큼

(threshold

- 특정 Rep nearest nei threshold 보다 크다 로 간주하고 Representative Point 가 포함된 anomalous po





2 HDoutliers Algorithm

HDoutliers Algorithm의 문제점

2

대표점 산출을 위한 중간 클러스터링 과정 추가

알고리즘 진행 과정

- 선제적으로 클러스터링 진행
- 각 클러스터에서 Representative Points 추출
- Representative Points들 간의 nearest neighbor distances들을 계산

HDoutliers Algorithm의 문제점

2

대표점 산출을 위한 중간 클러스터링 과정 추가

알고리즘 진행 과정

- 선제적으로 클러스터링 진행
- 각 클러스터에서 Representative Points 추출
- Representative Points들 간의 nearest neighbor distances들을 계산

HDoutliers Algorithm의 문제점

1

대표점 산출을 위한 중간 클러스터링 과정 추가

파생되는 문제점

(1) 클러스터링과 Representative Points, 그리고 threshold를 계산하는 과정에서 데이터의 density 특징을 완전히 무시

(2) 차원이 늘어날 때마다 연산량이 지수적으로 증가

필요한 가정 : Isolated Anomalies

- “정상치와 이상치 간의 거리는 멀다” 가정이 충족되어야 함

- 해당 가정 하에, 이상치는 nearest neighbor distance가 매우 큼
(threshold에 관해서는 추후 설명)

- 특정 Representative Point가 nearest neighbor distance가 threshold 보다 크다면 anomalous point로 간주하고, 해당Representative Point가 포함된 클러스터 전체가 anomalous points로 판별

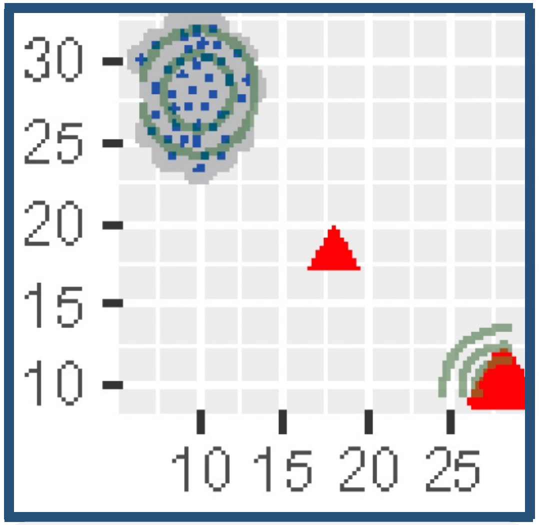
HDoutliers Algorithm의 문제점

2

대표점 산출을 위한 중간 클러스터링 과정 추가

예시 설명

- 좌상단, 우하단 클러스터에 각 각 1000개의 관측치 할당
- 이상치 클러스터에 1개 할당
- 앞서 설명한 알고리즘을 사용하면 우하단 클러스터는 1000개의 관측치가 있음에도 이상치로 판별됨 (Bimodal)



정상 : 좌상단, 우하단
이상치 : 중앙

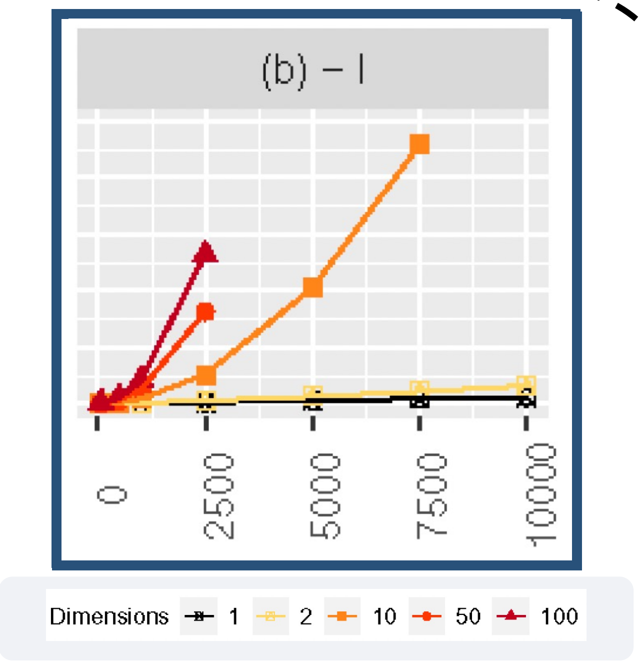
HDoutliers Algorithm의 문제점

2

대표점 산출을 위한 중간 클러스터링 과정 추가

예시 설명

- X축 : Sample Size
- Y축 : Running Time (miliseconds)
- 같은 Sample Size에서 차원이 높을수록 연산시간이 지수적으로 높아짐을 확인 가능



HDoutliers Algorithm의 문제점

3

Anomalous Threshold 계산 과정에서

FN rate을 증가시키는 경향이 있음

Anomalous Threshold

- Weissman's Spacing Theorem 사용

Weissman's Spacing Theorem :

- 어떤 분포 F로부터 독립적으로 X_i's들을 추출하고, $X_{i:n}$을 순서통계량으로 정의 ($\max(X_i) = X_{1:n}$, $\min(X_i) = X_{n:n}$, $\mathrm{i} = 1, 2, \dots, n$)
- $D_{j,n} = X_{j:n} - X_{j+1:n}$, $\mathrm{j} = 1, 2, \dots, k$ 로 정의
- 이 때, F가 maximum domain of attraction of the Gumbel Distribution에 속한다면, $D_{j,n}$'s들은 근사적으로 독립이며, 기대값이 $E(D_{j,n}) \propto j^{-1}$인 Exponential Distribution을 따름

2 HDoutliers Algorithm

HDoutliers Algorithm의 문제점

3

Anomalous Threshold 계산 과정에서

FN rate을 증가시키는 경향이 있음

Anomalous Threshold

- Weissman's Spacing Theorem 사용

Weissman's Spacing Theorem :

- 어떤 분포 F로부터 독립적으로 X_i's들을 추출하고, $X_{i:n}$을 순서통계량으로 정의 ($\max(X_i) = X_{1:n}$, $\min(X_i) = X_{n:n}$, $i = 1, 2, \ldots, n$)
- $D_{j,n} = X_{j:n} - X_{j+1:n}$, $j = 1, 2, \ldots, k$ 로 정의
- 이 때, F가 maximum domain of attraction of the Gumbel Distribution에 속한다면, $D_{j,n}$'s들은 근사적으로 독립이며, 기대값이 $E(D_{j,n}) \propto i^{-1}$인 Exponential Distribution을 따름

Maximum Domain of Attraction (MDA)

다음과 같은 Cumulative Distribution Function F를 MDA라고 정의함

→ There exists sequences of constants, C_n and D_n with $C_n > 0$ for all n, such that $\lim_{n \to \infty} P(\frac{M_n - D_n}{C_n} \leq x) = \lim_{n \to \infty} P(M_n - D_n \leq C_n\, x)$

$$\vdots$$

$$\lim_{n \to \infty} F^n(C_n\, x + D_n) = H(\mathrm{x}), \text{where } M_n = \max(X_\mathrm{i})$$

Maximum Domain of Attraction (MDA)

다음과 같은 Cumulative Distribution Function F를 MDA라고 정의함

- Fisher-Tippet Theorem에 의해 $H(x)$는 Gumbel, Frechet, Weibull Family 하나가 됨

- 즉, maximum을 적당한 C_n, D_n로 scaling해주면 위 3개 중 하나로 수렴함.

- Fisher-Tippet Theorem에 의해 $H(x)$는 Gumbel, Frechet, Weibull Family 하나가 됨

- 즉, maximum을 적당한 C_n, D_n로 scaling해주면 위 3개 중 하나로 수렴함.

식 유도

Assumption : $D_{i,n} \sim Exp(\lambda i)$

→ $f\left(D_{i,n} \,|\lambda\right) = (\lambda i)\exp[-(\lambda i)D_{i,n}]$
→ $l(\lambda|D_{1,n}, \dots, D_{k,n}) = (\mathrm{k}-1)\log(\lambda) + \sum_{i=2}^{k}\log(i) - \sum_{i=2}^{k}(\lambda i)D_{i,n}$
→ $\frac{\partial}{\partial\lambda} l(\lambda|D_{1,n}, \dots, D_{k,n}) = \frac{k-1}{\lambda} - \sum_{i=2}^{k}(i)D_{i,n}$
→ By the MLE,

$$\widehat{\lambda^{-1}} = \frac{1}{k-1}\left(\sum_{i=2}^{k}(i)D_{i,n}\right)$$

Then, let t be the anomalous threshold, *i.e.,* $P\left(D_{1,n} \leq t\right) = 1-\alpha$
→ $P\left(D_{1,n} \leq t\right) = 1 - \exp(-\lambda t) = 1-\alpha$
→ $\exp(-\lambda t) = \alpha$
→ $t = \frac{1}{\lambda}\log(\frac{1}{\alpha})$
→ By the Invariance property of MLE,

$$\hat{t} = \{\frac{1}{k-1}\left(\sum_{i=2}^{k}(i)D_{i,n}\right)\}\log(\frac{1}{\alpha})$$

식 유도

Assumption : $D_{i,n} \sim Exp(\lambda i)$

→ $f\left(D_{i,n} \mid \lambda\right) = (\lambda i)\exp[-(\lambda i)D_{i,n}]$
→ $l(\lambda|D_{1,n}, \ldots, D_{k,n}) = (\mathrm{k}-1)\log(\lambda) + \sum_{i=2}^{k}\log(i) - \sum_{i=2}^{k}(\lambda i)D_{i,n}$
→ $\frac{\partial}{\partial\lambda} l(\lambda|D_{1,n}, \ldots, D_{k,n}) = \frac{k-1}{\lambda} - \sum_{i=2}^{k}(i)D_{i,n}$
→ By the MLE,

$$\widehat{\lambda^{-1}} = \frac{1}{k-1}\left(\sum_{i=2}^{k}(i)D_{i,n}\right)$$

Then, let t be the anomalous threshold, *i.e.*, $P\left(D_{1,n} \le t\right) = 1-\alpha$
→ $P\left(D_{1,n} \le t\right) = 1 - \exp(-\lambda t) = 1-\alpha$
→ $\exp(-\lambda t) = \alpha$
→ $t = \frac{1}{\lambda}\log\left(\frac{1}{\alpha}\right)$
→ By the Invariance property of MLE,

$$\hat{t} = \left\{\frac{1}{k-1}\left(\sum_{i=2}^{k}(i)D_{i,n}\right)\right\}\log\left(\frac{1}{\alpha}\right)$$

HDoutliers Algorithm의 문제점

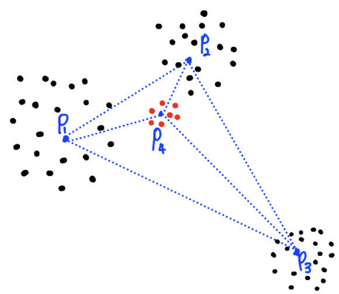
3

Anomalous Threshold 계산 과정에서

FN rate을 증가시키는 경향이 있음

Anomalous Threshold

(EX)



representative points

Nearest Neighbor Distance for each P_i :

$P_1 : P_4 \Rightarrow \text{distance} = 2 = X_1$ $\quad X_{1:4} = 7 = max(X_1, X_2, X_3, X_4)$

$P_2 : P_4 \Rightarrow \text{distance} = 1 = X_2$ $\quad \Rightarrow X_{2:4} = 2$

$P_3 : P_4 \Rightarrow \text{distance} = 7 = X_3$ $\quad X_{3:4} = 1$

$P_4 : P_2 \Rightarrow \text{distance} = 1 = X_4$ $\quad X_{4:4} = 1 = min(X_1, X_2, X_3, X_4)$

$\Rightarrow D_{1,4} = 0 \,,\; D_{2,4} = 1 \,,\; D_{3,4} = 5$

Assuming $\alpha = 0.05$

$\Rightarrow \hat{t} = \frac{1}{3-1}\left(\sum_{i=1}^{3} i \cdot D_{i,n}\right) log\left(\frac{1}{0.05}\right) \,, \quad \sum_{i=1}^{3} i \cdot D_{i,n} = 0 + 2 + 15 = 17$

$= 11.06$

$\Rightarrow P(P_i \le \hat{t})$; if $P_i > \hat{t} = 11.06$, P_i is anomaly, for fixed i

$\therefore$ By the algorithm, every point is normal, indicating the increase in the False Negative

HDoutliers를 사용하게 되면 threshold가 커지면서 FN이 증가

2 HDoutliers Algorithm

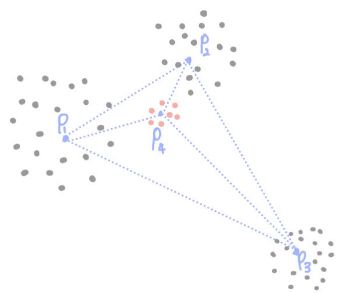
HDoutliers Algorithm의 문제점

3

Anomalous Threshold 계산 과정에서
FN rate을 증가시키는 경향이 있음

Anomalous Threshold

(EX)



$\overbrace{P_i}^{\text{representative points}}$

Nearest Neighbor Distance for each P_i :

$$\left.\begin{array}{l} P_1 : P_4 \Rightarrow \text{distance} = 2 = X_1 \\ P_2 : P_4 \Rightarrow \text{distance} = 1 = X_2 \\ P_3 : P_4 \Rightarrow \text{distance} = 7 = X_3 \\ P_4 : P_2 \Rightarrow \text{distance} = 1 = X_4 \end{array}\right\} \Rightarrow \begin{array}{l} X_{1:4} = 7 = \max(X_1, X_2, X_3, X_4) \\ X_{2:4} = 2 \\ X_{3:4} = 1 \\ X_{4:4} = 1 = \min(X_1, X_2, X_3, X_4) \end{array}$$

$$\Rightarrow D_{1,4} = 0,\ D_{2,4} = 1,\ D_{3,4} = 5$$

Assuming $\alpha = 0.05$

$$\Rightarrow \hat{t} = \frac{1}{3-1}\left(\sum_{i=1}^{3} i \cdot D_{i,n}\right) \log\left(\frac{1}{0.05}\right),\quad \sum_{i=1}^{3} i \cdot D_{i,n} = 0 + 2 + 15 = 17$$

$$= 11.06$$

$\Rightarrow P(P_i \le \hat{t})$; if $P_i > \hat{t} = 11.06$, P_i is anomaly, for fixed i

$\therefore$ By the algorithm, every point is normal, indicating the increase in the False Negative

HDoutliers를 사용하게 되면 threshold가 커지면서 FN이 증가

3

STRAY ALGORITHM

알고리즘의 특징

(1) 빠른 연산으로 실시간 적용가능

(2) KNN을 활용해 masking problem에 효과적으로 대응

(3) Multimodal distribution을 따르는 데이터에도 효과적

(4) 이진분류는 물론 anomalous score를 함께 제공

알고리즘의 특징

(1) 빠른 연산으로 실시간 적용가능

(2) KNN을 활용해 masking problem에 효과적으로 대응

(3) Multimodal distribution을 따르는 데이터에도 효과적

(4) 이진분류는 물론 anomalous score를 함께 제공

HDoutlier로부터의 개선점

(1) NN distance → KNN distance with the maximum gap

(2) 중간 clustering단계가 생략되어 연산속도 증가

(3) 분류 only → 분류 및 anomalous score 제공

알고리즘 과정

1

정규화 : min-max (수치형), correspondence (범주형)

2

각 점들에 대해 KNN with the maximum gap 계산 및 순서 정렬

3

하위 50%의 순서통계량에 한해서 Spacing Theorem으로 threshold 계산

4

- 상위 50%의 순서통계량에 오름차순으로 threshold와 비교
- $\hat{t}$보다 작으면 정상으로 분류, $\hat{t}$보다 크면 해당 통계량 포함하여 위의 나머지를 모두 이상치로 분류

알고리즘 과정

1

정규화 : min-max (수치형), correspondence (범주형)

Correspondence Analysis

- 다변량 통계기법 중 하나로, 개념적으로 PCA와 유사
- 단, PCA는 수치형 자료에, CA는 범주형 자료에 사용

모두 0~1 사이의 값을 가지게 정규화

알고리즘 과정

1

정규화 : min-max (수치형), correspondence (범주형)

Correspondence Analysis

- 다변량 통계기법 중 하나로, 개념적으로 PCA와 유사
- 단, PCA는 수치형 자료에, CA는 범주형 자료에 사용

모두 0~1 사이의 값을 가지게 정규화

알고리즘 과정

2

각 점들에 대해 KNN with the maximum gap 계산 및 순서 정렬

KNN with the Maximum Gap

- K개의 가장 가까운 점을 구하고, 이 거리들의 차이를 계산
- 거리들의 차이값 중 가장 차이가 많이 나는 값으로 정의

Masking problem에 효과적으로 대처함과 동시에
HDoutliers의 문제점이었던 높은 threshold에도 효과적으로 대응

3 STRAY ALGORITHM

알고리즘 과정

2

각 점들에 대해 KNN with the maximum gap 계산 및 순서 정렬

KNN with the Maximum Gap

- K개의 가장 가까운 점을 구하고, 이 거리들의 차이를 계산
- 거리들의 차이값 중 가장 차이가 많이 나는 값으로 정의

Masking problem에 효과적으로 대처함과 동시에
HDoutliers의 문제점이었던 높은 연산시간에도 효과적으로 대응
(real-time data에도 적용가능)

3 STRAY ALGORITHM

알고리즘 과정

3

하위 50%의 순서통계량에 한해서 Spacing Theorem으로 threshold 계산

- HDoutliers의 문제점이었던 높은 threshold에 효과적으로 대응
- Threshold를 적절하게 산정하면서 FN rate을 낮춤

4

- 상위 50%의 순서통계량에 오름차순으로 threshold와 비교
- $\hat{t}$보다 작으면 정상으로 분류, $\hat{t}$보다 크면 해당 통계량 포함하여 위에 나머지를 모두 이상치로 분류

- Multimodal Distribution에도 효과적으로 작동
- Anomalous scores도 제공하여 해석 가능

알고리즘 과정

3

하위 50%의 순서통계량에 한해서 Spacing Theorem으로 threshold 계산

- HDoutliers의 문제점이었던 높은 threshold에 효과적으로 대응
- Threshold를 적절하게 산정하면서 FN rate을 낮춤

4

- 상위 50%의 순서통계량에 오름차순으로 threshold와 비교
- $\hat{t}$보다 작으면 정상으로 분류, $\hat{t}$보다 크면 해당 통계량 포함하여 위에 나머지를 모두 이상치로 분류

- Multimodal Distribution에도 효과적으로 작동
- Anomalous scores도 제공하여 해석 가능

알고리즘 과정

3

하위 50%의 순서통계량에 한해서 Spacing Theorem으로 threshold 계산

- HDoutliers의 문제점이었던 높은 threshold에 효과적으로 대응
- Threshold를 적절하게 산정하면서 FN rate을 낮춤

4

- 상위 50%의 순서통계량에 오름차순으로 threshold와 비교
- $\hat{t}$보다 작으면 정상으로 분류, $\hat{t}$보다 크면 해당 통계량 포함하여 위에 나머지를 모두 이상치로 분류

- Multimodal Distribution에도 효과적으로 작동
- Anomalous scores도 제공하여 해석 가능

알고리즘 과정

3

하위 50%의 순서통계량에 한해서 Spacing Theorem으로 threshold 계산

- HDoutliers의 문제점이었던 높은 threshold에 효과적으로 대응
- Threshold를 적절하게 산정하면서 FN rate을 낮춤

4

- 상위 50%의 순서통계량에 오름차순으로 threshold와 비교
- $\hat{t}$보다 작으면 정상으로 분류, $\hat{t}$보다 크면 해당 통계량 포함하여 위에 나머지를 모두 이상치로 분류

- Multimodal Distribution에도 효과적으로 작동
- Anomalous scores도 제공하여 해석 가능

알고리즘 과정

3

하위 50%의 순서통계량에 한해서 Spacing Theorem으로 threshold 계산

- HDoutliers의 문제점이었던 높은 threshold에 효과적으로 대응
- Threshold를 적절하게 산정하면서 FN rate을 낮춤

4

- 상위 50%의 순서통계량에 오름차순으로 threshold와 비교
- $\hat{t}$보다 작으면 정상으로 분류, $\hat{t}$보다 크면 해당 통계량 포함하여 위에 나머지를 모두 이상치로 분류

- Multimodal Distribution에도 효과적으로 작동
- Anomalous scores도 제공하여 해석 가능

감사합니다