논문: https://doi.org/10.1007/s40745-021-00362-9

A Comprehensive Survey of Anomaly Detection Algorithms

이상 탐지 알고리즘에 대한 종합적인 조사

박세현
LAMDA Lab in SKKU

이상 탐지의 정의

- 이상 징후의 고전적 정의
  1. 해당 관측치가 발생한 표본의 다른 구성원과 현저하게 다르게 나타나는 것
  2. 다른 관측치와 너무 많이 달라서 다른 메커니즘에 의해 생성되었다는 의심을 불러일으키는 관측치
  3. 나머지 데이터 집합과 일치하지 않는 것으로 보이는 관찰(또는 관찰의 하위 집합)
  4. 해당 지역 이웃의 밀도에 비해 낮은 지역 밀도에 있는 점
  5. 데이터 집합의 클러스터에 속하지 않거나 다른 클러스터보다 현저히 작은 클러스터로 존재하는 점
  6. 점의 밀도가 인근 고밀도 패턴 클러스터 보다 상대적으로 낮거나 자체 밀도가 인근 저밀도 패턴 규칙성 보다 상대적으로 높은 상황
  7. 잘 정의된 정상 동작의 개념에 부합하지 않는 데이터의 패턴
  8. 크게 벗어나 정상 데이터와 일치하지 않는 데이터 레코드 또는 인스턴스

이상 탐지의 정의

- 정의들은 서로 다르지만 유사한 개념
  -> 나머지 데이터와 일치하지 않는 데이터 포인트를 찾는 작업

- Anomaly detection =
  Outlier detection, Novelty detection, Abnormality detection

이상의 유형

- Point anomaly = Global anomaly
  : 데이터 포인트가 나머지 데이터 포인트와 크게 차이가 나는 이상

- Group anomaly
  : 데이터 포인트의 모음으로 나머지 데이터 포인트와 비교한 이상

- Local anomaly
  : 데이터 포인트 인근에서 발생한 이상

- Collective anomaly
  : 전체 데이터셋과 비교하여 이상이 있는 지점의 집합적 이상
  (그룹 내에 인스턴스는 이상치가 아닐 수 있지만, 이상 지역에 존재하기때문에 이상치로 간주될 수 있음)

이상 탐지 응용 분야

- 침입 탐지 – 컴퓨터나 네트워크 시스템의 비정상적인 활동
- 고장 진단(탐지) – 기계 장치의 결함
- 의료 – 환자의 특이한 건강 상태
- 사기 탐지 – 신용카드 거래 또는 보험 청구와 관련된 사기
- 텍스트로부터 신규성 탐지 – 새로운 텍스트, 뉴스 또는 문서의 집합

평가 지표

- Precision at n (P@n) $\mathrm{P@n} = \frac{\{|a \in \mathcal{A} | rank(a) \leq n|\}}{n}$
- Average precision $\mathrm{AP} = \frac{1}{n} \sum_{a \in \mathcal{A}} P@rank(a)$
- ROC AUC

이상 탐지 알고리즘 범주

- 7개 범주, 52개 알고리즘
  - 통계
  - 밀도
  - 거리
  - 클러스터
  - 격리
  - 앙상블
  - 서브스페이스

Table 4 Anomaly detection algorithms

Category	Anomaly detection algorithm
Anomaly detection algorithms based on statistic model	Grabbs'test, Dixon test, Rosner's test, Student's t-test, Hostelling t^2-test, χ^2-statistics test, box plots, HBOS
Anomaly detection algorithms based on density	LOF, COF, LoOP, LOCI, RDF, INFLO, ROF, FastLOF, DWOF, SimplifiedLOF, LiNearN, GLOSH, SPAD, SPAD+
Anomaly detection algorithms based on distance	k-NN, kth-NN, RBRP, ABOD, GPA, LDOF, Sp, AntiHub
Anomaly detection algorithms based on clustering	OFP, FindOut, FindCBLOF, CBOD
Anomaly detection algorithms based on isolation	iForest, SciForest, HS-Tree, ReMass-iForest, iNNE, LeSiNN, LSHiForest, usfAD
Anomaly detection algorithms based on ensemble	LODA, DCSO, LSCP
Anomaly detection algorithms based on subspace	SOD, LSOF, HighDOD, COP, HiCS, CMI, Zero++

통계 기반

- 저확률 구간을 이상치로 간주
- 장점
  - 히스토그램 기반 방법은 다른 이상 탐지 알고리즘에 비해 매우 단순하고 직관적
- 단점
  - 대부분의 방법은 단변량 데이터에만 적용 -> 다차원 데이터 셋을 처리할 때 계산 비용 매우 높음
  - 히스토그램 기반 방법은 다중 특징에 의존하는 이상을 포착하지 못함
  - 커널 기반 방법은 계산 비용이 매우 높고 매개변수에 민감

통계 기반

Table 5 Analysis of different statistic tests

Name	Function equation	Analysis
Grubbs's test [4]	$G = \frac{\max_{i=1,\ldots,n} \lvert x_i - \bar{\mathcal{X}} \rvert}{s}$	• Also known as maximum normalized residual test • G greater than critical value is considered as an anomaly • Easy to implement • For univariate data set only • One anomaly at a time • Where $\bar{\mathcal{X}}$ and s are sample mean and standard deviation, respectively
Dixon test [31]	$Q = \frac{x_n - x_{n-1}}{x_n - x_1}$	• Assume—only one outlier is present • Simple to implement • Applicable only on small datasets • First need to arrange data in ascending order and then compute the Dixon test on data
Rosner's test [21]	$R_{i+1} = \frac{\lvert x_i - \bar{\mathcal{X}} \rvert}{s}$	• where $\bar{\mathcal{X}}$ and s are as defined above • The data follow a normal distribution and the anomalies are employed from a different distribution • This test is good for larger data sets • Not simple as Dixon test • Must have knowledge of number of anomalies
Student's t-test [23,24]	$t = \frac{Z}{\sigma} = \frac{\bar{\mathcal{X}} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$	• Where $\bar{\mathcal{X}}$ is as defined above, μ and $\hat{\sigma}$ are mean and standard deviation of population, respectively and n is total number of samples • Normal samples are compared to test instance
Hostelling t^2-test [22]	$T^2 = n(\bar{\mathcal{X}} - \mu)' S^{-1} (\bar{\mathcal{X}} - \mu)$	• Where $\bar{\mathcal{X}}$, n and μ are as defined above. S is a covariance matrix
χ^2-statistics [26]	$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\mu}$	• μ is defined as above • *Assumption*—normal data has a multidimensional normal distribution • Data point with higher χ^2 value is considered as an anomaly

밀도 기반

- 주변 밀도 대비 낮은 포인트 탐지
- 장점
  - 매우 직관적 -> 널리 사용됨
  - 통계 및 거리 기반 알고리즘 보다 성능 우수
- 단점
  - 쌍별 거리를 계산 -> 계산 비용 높음
  - 대규모 및 고차원 데이터셋에는 적합하지 않음
  - 가장 가까운 이웃의 크기(k)와 같은 매개변수에 민감

밀도 기반

Table 6 Time complexity of anomaly detection algorithms based on density

Methods	Time complexity	
	Training stage	Testing stage
LOF	–	$O(n^2d)$
COF	–	$O(n^2d)$
LoOP	–	$O(n^2d)$
LOCI	–	$O(n^3)$
RDF	–	$O(n^2d)$
INFLO	–	$O(n^2d)$
ROF	–	$O(n^2d)$
FastLOF	–	$O(n^2d)$
SimplifiedLOF	–	$O(n^2d)$
LiNearN	$O(t(\psi+\Psi)\psi d))$	$O(nt\psi d)$
SPAD	–	$O(nd)$
SPAD+	$O(nth+t\psi d)$	$O(t(h+d))$

거리 기반

- 최근접 이웃까지 거리로 이상치 측정
- 장점
  - 구현이 쉽고 직관적
  - 데이터 분포에 독립적
  - 인덱싱 구조 사용 하면 시간 복잡도 O(n log(n))
- 단점
  - 일반적으로 높은 시간 복잡도
  - 인덱싱 방안은 고차원 데이터 세트에서 작동 불가

거리 기반

Table 7 Time complexity of anomaly detection algorithms based on distance. #m is total number of cells

Methods	Time complexity
kNN	$O(n^2 d)$
kth-NN	$O(n^2 d)$
RBRP	$O(n^2 d)$
ABOD	$O(n^3 d)$
FastABOD	$O(n^2 + nk^2)$
GPA	$O(kn^2 + m)$
LDOF	$O(n^2 d)$
Sp	$O(nd\psi)$
AntiHub	$O(n^2 d)$

클러스터 기반

- 클러스터에 속하지 않거나 작은 클러스터
- 장점
  - 비지도 학습 환경에서 쉽게 적용가능
  - 클러스터링 알고리즘을 단순히 교체 가능 -> 복잡하고 다양한 데이터 유형과 호환
- 단점
  - 성능이 클러스터링 알고리즘에 크게 의존
  - 높은 시간 복잡도
  - 매개변수에 민감
  - 이진 점수를 사용 -> 강한 이상, 약한 이상 구분 불가

클러스터 기반

Table 8 Time complexity of anomaly detection algorithms based on clustering

	OFP	FindOut	FindCBLOF	CBOD
Time complexity	$O(n^2 d)$	$O(Tdn\log_2(n))$	$O(n^2 d)$	$O(n^2 d)$

격리 기반

- 무작위 분할로 포인트 격리 길이 활용
- 장점
  - 상대적으로 낮은 시간 복잡도와 높은 정확도
  - 높은 확장성
  - 글로벌 및 로컬 이상값이 있는 데이터에 적합
  - SCiForest는 iForest, LOF와 같은 다른 방법보다 로컬 및 글로벌 클러스터링 이상값 탐지에 높은 정확도
- 단점
  - 트리 기반 방법은 로컬 이상에 부적합
  - HS-Tree는 스트리밍 데이터에만 적용 가능
  - iForest의 추가 단점 존재

격리 기반

Table 9 Time complexity of different anomaly detection algorithms based on isolation

Methods	Time complexity Training stage	Testing stage
iForest	$O(t\psi \log(\psi))$	$O(nt \log(\psi))$
SCiForest	$O((t\tau\psi(q\psi + \log(\psi) + \psi))$	$O(qnt\psi)$
HS-Tree	–	$O(t(h + \psi))$
ReMass-iForest	$O(t\psi \log(\psi))$	$O(nt \log(\psi))$
iNNE	$O(t\psi^2 d)$	$O(ntd\psi)$
LeSiNN	$O(\psi td)$	$O(n\psi td)$
LSHiForest	$O(t\psi \log(\psi)d)$	$O(nt \log(\psi)d)$
usfAD	$O(nth + t\psi d)$	$O(t(h + d))$

앙상블 기반

- 다양한 탐지기의 오류 다양성 결합
- 장점
  - 일반적으로 매우 안정적이며 우수한 성능 발휘
  - 이상치 분석에 유용
- 단점
  - 개발된 방법이 매우 적음
  - 상대적으로 높은 시간 복잡도

앙상블 기반

Table 10 Time complexity of anomaly detection algorithms based on ensemble

Methods	LODA	DSCO	LSCP
Time complexity	$O(nkd^{-\frac{1}{2}})$	$O(nd + n\log(n))$	$O(nd + n\log(n))$

서브스페이스 기반

- 일부 특성 하위공간에서만 드러나는 이상 탐지
- 장점
  - 숨겨진 이상 현상을 탐지하는데 우수
- 단점
  - 높은 시간 복잡도
  - Zero++는 범주형 데이터 세트에만 적용 가능
  - 관련 없는 속성에 민감

서브스페이스 기반

Table 11 Time complexity of anomaly detection algorithms based on subspace

Methods	Time complexity
SOD	$O(n^3 d)$
LSOF	$O(n^2 d)$
HighDOD	$O((x+n)Ndim(S))$
COP	$O(n^2 d^3)$
HiCS	$O(n^2 d)$
CMI	$O(n^2 d)$
Zero++	$O(ntq + dtq)$

종합 통찰

Table 12 Comparison of different anomaly detection algorithms

Category	Methods	Equation	Time Complexity	Scalability
Based on density	LOF	$\mathrm{LOF}(x)=\frac{\sum_{y\in N^k(x)} lrd(y)}{\lvert N^k(x)\rvert\times lrd(x)}$	High	×
	COF	$\mathrm{COF}(x)=\frac{\lvert N_k(x)\rvert\cdot dist_{N_k(x)}(x)}{\sum_{y\in N_k(x)}\cdot dist_{N_k(y)}(y)}$	High	×
	LoOP	$\mathrm{LoOP}(x)=\max\left\{0,\mathrm{erf}\left(\frac{PLOF_{\lambda,S}(x)}{nPLOF\cdot\sqrt{2}}\right)\right\}$	High	×
	LOCI	$\mathrm{MDEF}(x_i,r,\alpha)=1-\frac{n(x_i,\alpha r)}{\hat{n}(x_i,r,\alpha)}$	High	×
	RDF	$\mathrm{RDF}(x,r)=\frac{DF_{nbr}(N_k(x),r)}{DF(N_k(x),r)}$	High	×
	INFLO	$\mathrm{INFLO}_k(x)=\frac{\sum_{y\in IS_k(x)} den(y)}{\lvert IS_k(x)\rvert\cdot den(x)}$	High	×
	ROF	$\mathrm{ROF}(x)=\sum_{i=1}^{R}\frac{\mathrm{ClusterSize}(x,r_{i-1})-1}{\mathrm{ClusterSize}(x,r_i)}$	High	×
	LiNearN	–	Low	✓
	SPAD	$\mathrm{SPAD}(x)=\sum_{i=1}^{d}\log\frac{\lvert H_i(x)\rvert+1}{n+b}$	Low	✓
	SPAD+	$\mathrm{SPAD+}(x)=\sum_{i=1}^{d}\log\frac{\lvert H_i(x)\rvert+1}{n+b}+\sum_{j=1}^{d}\log\frac{\lvert H_j(x')\rvert+1}{n+b}$	Low	✓
Based on distance	kNN	$k\text{-NN}=\sum_{y\in kNN(x)} dist(x,y)$	High	×
	kth-NN	$k\text{th-NN}(x):=dist_k(x;X)$	High	×
	RBRP	–	High	×
	ABOD	–	High	×
	FastABOD	–	High	×
	GPA	–	High	×

Table 12 continued

Category	Methods	Equation	Time Complexity	Scalability
	LDOF	$\mathrm{LDOF}(x)=\frac{\bar{d}_x}{D_x}$	High	×
	Sp	$\mathrm{Sp}(x)=\min_{y\in\mathcal{S}} dist(x,y)$	Low	✓
	AntiHub	–	High	×
Based on clustering	OFP	–	High	×
	FindOut	–	Low	✓
	FindCBLOF	–	High	×
	CBOD	–	High	×
Based on isolation	iForest	$\mathrm{iForest}(x)=\frac{1}{t}\sum_{i=1}^{t} l_i(x)$	Low	✓
	SCiForest	$\mathrm{SCiForest}(x)=\frac{1}{t}\sum_{i=1}^{t} l_i(x)$	Low	✓
	HS-Tree	$\text{HS-Tree}(x)=\frac{1}{t}\sum_{i=1}^{t} m_i(x)$	Low	✓
	ReMass-iForest	$\text{ReMass-iForest}(x)=\frac{1}{t}\sum_{i=1}^{t} s_i(x)$	Low	✓
	iNNE	$\mathrm{iNNE}(x)=\frac{1}{t}\sum_{i=1}^{t} I_i(x)$	Low	✓
	LeSiNN	$\mathrm{LeSiNN}(x)=\frac{1}{t}\sum_{i=1}^{t}\min_{y\in\mathcal{S}} dist(x,y)$	Low	✓
	LSHiForest	–	Low	✓
	usfAD	$\mathrm{usfAD}(x)=\frac{1}{t}\sum_{i=1}^{t} l_i(x)$	Low	✓

Table 12 continued

Category	Methods	Equation	Time Complexity	Scalability
Based on ensemble	LODA	–	Low	✓
	DSCO	–	Low	✓
	LSCP	–	Low	✓
Based on subspace	SOD	$\mathrm{SOD}_{R(x)}(x)=\frac{dist(y,\mathcal{H}(R(x)))}{\lVert v^{R(x)}\rVert_1}$	High	×
	LSOF	$\mathrm{LSOF}(x)=\frac{1}{\lvert N_{minPts}(x)\rvert}\sum_{y\in N_{minPts}(x)}\frac{lsrd(y)}{lsrd(x)}$	High	×
	COP	$\mathrm{COP}(x,\psi)=norm(1-\cos(x),\psi)$	High	×
	CMI	$\mathrm{CMI}(x)=-\sum_{i=1}^{n-1}\left(x_{i+1}-x_i\right)\frac{i}{n}\log\frac{i}{n}$	High	×
	Zero++	$\mathrm{Zero}(x,\mathcal{D}\vert S)=\sum_{i=1}^{\psi}\sum_{S\in\mathcal{S}} I(P_S(x\vert\mathcal{D})=0)$	Low	✓

종합 통찰

- 성능 속도
  - Isolation Forest 계열이 가장 빠르고 확장성이 높음
  - SCiForest는 국소/군집 이상치까지 검출력을 개선

- 데이터 특성 별 추천
  - 대규모/고차원: Isolation Forest, LSHiForest
  - 국소 밀집 이상: SCiForest, ReMass-iForest, 밀도 기반(LOCI 등)
  - 해석 가능성 중시: 통계/클러스터/서브스페이스 기반(box-plot, COP 등)