The Korean Journal of Applied Statistics
2021, Vol. 34, No. 3, 507–521

DOI: https://doi.org/10.5351/KJAS.2021.34.3.507
Print ISSN1225-066X / Online ISSN 2383-5818

A survey on unsupervised subspace outlier detection methods for high dimensional data

Jaehyeong Ahn[a], Sunghoon Kwon[1,a]

[a]Department of Applied Statistics, Konkuk University

Abstract

Detecting outliers among high-dimensional data encounters a challenging problem of screening the variables since relevant information is often contained in only a few of the variables. Otherwise, when a number of irrelevant variables are included in the data, the distances between all observations tend to become similar which leads to making the degree of outlierness of all observations alike. The subspace outlier detection method overcomes the problem by measuring the degree of outlierness of the observation based on the relevant subsets of the entire variables. In this paper, we survey recent subspace outlier detection techniques, classifying them into three major types according to the subspace selection method. And we summarize the techniques of each type based on how to select the relevant subspaces and how to measure the degree of outlierness. In addition, we introduce some computing tools for implementing the subspace outlier detection techniques and present results from the simulation study and real data analysis.

Keywords: outlier detection, high-dimensional data, subspace outlier detection

1. 서론

이상치 탐지(outlier detection)는 자료에 포함된 이상치를 탐지하기 위한 분석 기법으로 신용카드 사기 탐지 (Fawcett과 Provost, 1997), 네트워크 침입 탐지 (Eskin 등, 2002), 의료 진단 (Penny와 Jolliffe, 2001) 등 희귀한 현상에 대한 정확한 식별이 요구되는 분야에서 주로 활용되고 있다. 정량적 정의는 각각의 활용 분야에서 관측치의 이상정도(degree of outlierness)를 어떻게 정의하는가에 따라 결정되며 보통은 '어떤 자료 안에서 다른 관측치들로부터 멀리 떨어져 있어 다른 매커니즘에 의해 생성되었다고 의심되는 관측치' (Hawkins, 1980), '다른 관측치들과 비교하여 현저히 다른 관측치' (Barnett과 Lewis, 1984)와 같이 정성적인 의미로 정의되는 경우가 많다. 이상치 탐지기법 중 이상치 여부(outlier label)를 사용하지 않는 방법을 비지도 이상치 탐지기법 (unsupervised outlier detection) 이라고 하며 이 중 전체 변수의 일부를 선별하여 이상치를 탐지하는 기법을 부분공간 이상치 탐지기법(subspace outlier detection)이라고 한다.

부분공간 이상치 탐지기법이 필요한 이유 중 하나는 다음과 같은 집중효과(concentration effect) 현상이다.

$$\mathrm{Var}\left(\frac{\|X\|}{E\|X\|}\right)\to_{d\to\infty} 0 \;\Rightarrow\; \frac{\max_{i\le n}\|x_i\|}{\min_{i\le n}\|x_i\|} \xrightarrow{p}_{d\to\infty} 1.$$

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.NRF-2020R1F1A1A01071036).

[1] Corresponding author: Department of Applied Statistics, Konkuk University, 120 Neungdong-ro Gwangjin-Gu, Seoul 05029, Korea. E-mail: shkwon0522@konkuk.ac.kr

Published 30 June 2021 / journal homepage: http://kjas.or.kr

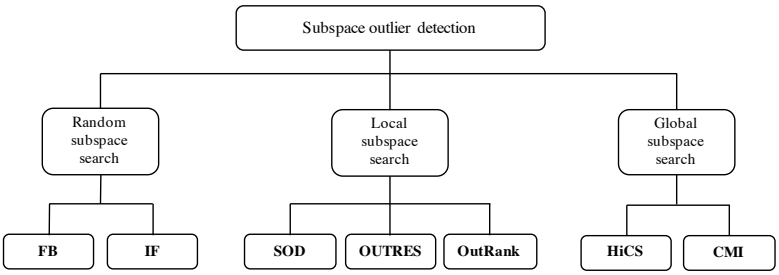

Figure 1: *Taxonomy of subspace outlier detection methods.*

위 식에서 $X \in \mathbb{R}^d$는 자료에 포함된 d개의 변수를 나타내는 확률벡터이고 x_i, $i \leq n$는 X에 대한 n개의 임의표본(random sample)이다. 자료에 포함된 변수의 개수가 증가하면 즉, 고차원 자료이면 모든 관측치 사이의 거리가 비슷해짐을 의미한다. 대부분의 비지도 이상치 탐지기법은 관측치 사이의 거리에 기반해 이상정도를 측정하기 때문에 집중효과가 나타날 경우 모든 관측치의 이상정도가 비슷해져 이상치와 정상치의 구분이 어려워지는 문제가 발생하게 된다. Beyer 등 (1999)는 자료에 포함된 변수가 동일 분포를 따르는 임의표본인 경우를 비롯한 다양한 상황에서 집중효과가 발생함을 증명하였다. Bennett 등 (1999)은 자료가 분리 가능한 클러스터(separable cluster) 구조를 가질 경우 서로 다른 클러스터에 속한 관측치 사이의 거리가 동일한 클러스터에 속한 관측치 사이의 거리보다 항상 크다는 것을 보임으로써 집중효과가 나타나지 않음을 입증하였다. Durrant과 Kabán (2009)은 변수들 사이에 강한 상관관계가 존재하여 내재적인 잠재구조(latent structure)를 형성하는 경우 집중효과가 나타나지 않음을 증명하였다. 또한 잠재적인 구조에 기여하지 않는 변수를 부적합한 변수(irrelevant variable)로 정의하여 전체 변수 중 부적합한 변수가 차지하는 비중이 높을수록 집중효과가 두드러짐을 경험적으로 보였다. Zimek 등 (2012)은 전체 변수 중 자료의 잠재구조에 기여하는 즉, 적합한 변수(relevant variable)의 비중이 높을수록 집중효과의 영향이 줄어들어 이상치와 정상치간 이상정도의 대비가 뚜렷해진다는 연구결과를 발표하였다.

부분공간 이상치 탐지기법은 클러스터 구조를 가지거나 강한 상관관계를 갖는 등 특정 기준을 만족하는 부분공간을 선정하여 관측치의 이상정도를 측정함으로써 고차원 자료에서의 집중효과를 극복한다. 본 논문은 대표적인 비지도 부분공간 이상치 탐지기법을 부분공간 선정 방식에 따라 임의적 부분공간 탐색(random subspace search), 지역적 부분공간 탐색(local subspace search), 전역적 부분공간 탐색(global subspace search)의 세가지 유형으로 분류하고 각 유형에 속한 방법론을 부분공간 선정기준과 이상정도 측정 방식에 따라 개괄한다. 또한 이러한 분류기준에 따라 방법론을 요약한 표를 제공하고 방법론을 적용할 수 있는 컴퓨팅 프로그램을 소개하여 부분공간 이상치 탐지기법에 대한 연구자들의 이해도와 접근성을 높이고자 한다. 마지막으로 부분공간 이상치 탐지기법의 필요성에 대한 이해를 돕기 위해 집중효과에 대한 간단한 가상실험을 수행하며 이상치가 존재하는 실제 자료에 각 방법론을 적용한 결과를 비교한다. Figure 1은 본 논문에서 다루는 부분공간 이상치 탐지기법을 요약하여 나타낸다.

2. 부분공간 이상치 탐지기법

먼저 기술에 필요한 기호를 몇 가지 정의하기로 한다. 임의의 집합 혹은 벡터 A에 대하여 $\text{Lower}_k(A)$는 A의 원소 중 k번째로 작은 원소, $\text{Upper}_k(A)$는 A의 원소 중 k번째로 큰 원소, $\text{Sum}(A)$는 A의 원소의 합, $\text{Product}(A)$는 A의 원소의 곱, $\text{Average}(A)$는 A의 원소의 평균, $|A|$는 A의 원소의 개수라고 하자. $X = (X_1, \ldots, X_d) \in \mathbb{R}^{n \times d}$를 d개의 변수 $X_j = (x_{1j}, \ldots, x_{nj})^T$와 n개의 관측치 $x_i = (x_{i1}, \ldots, x_{id})^T$로 구성된 자료라고 하자. 전체 변수의 인덱스(index) 집합 $\mathcal{D} = \{1, \ldots, d\}$를 전체공간, 부분집합 $\mathcal{S} \subset \mathcal{D}$를 부분공간, $|\mathcal{S}|$를 $\mathcal{S}$의 차원이라고 하자. $\text{Dist}_{\mathcal{S}}(x_s, x_t)$는 두 관측치 x_s와 x_t에 대하여 주어진 부분공간 $\mathcal{S}$에 포함된 변수만을 이용해 계산한 부분공간 거리(distance

on subspace)라고 하자. 예를 들어 유클리디안 거리를 사용하는 경우 $\mathrm{Dist}_{\mathcal{S}}(x_s, x_t) = \mathrm{Sum}(\{|x_{sj} - x_{tj}|^2 : j \in \mathcal{S}\})^{1/2}$ 이다.

2.1. 임의적 부분공간 탐색(Random subspace search)

임의적 부분공간 탐색기법은 임의로 선정한 부분공간에서 관측치의 부분공간 이상정도(subspace degree of outlierness)를 측정하는 기법이다. 이 기법은 부분공간 선정의 임의성을 보완하기 위하여 여러 개의 부분공간에서 계산된 부분공간 이상정도를 평균과 같은 결합함수(combination function)로 요약하여 최종 이상정도를 정의한다.

2.1.1. Feature bagging (FB)

FB (Lazarevic과 Kumar, 2005)는 가장 단순한 임의적 부분공간 탐색기법으로 적당한 차원의 부분공간을 임의로 선정하고 선정된 부분공간에 대해 기존의 이상치 탐지기법(classical outlier detection)을 적용하여 부분공간 이상정도를 계산하는 과정을 반복한 후 계산된 부분공간 이상정도를 누적합(cumulative sum)등으로 요약하는 기법이다. **FB**에서 관측치 x_i의 이상정도를 계산하는 절차는 다음과 같다.

(A) 사용자 정의 요소를 결정한다.

- COD: 부분공간에서 사용할 이상치 탐지기법
- m: 선정할 부분공간의 개수

(B) 부분공간 집합(set of subspaces)을 선정한다.

$$\mathrm{SS} = \{\mathcal{S}^k : k \le m\},$$

- $\mathcal{S}^k$: $d/2 \le |\mathcal{S}^k| \le d-1$을 만족하는 임의의 부분공간

(C) 관측치의 이상정도를 계산한다.

$$\mathrm{FB}(x_i) = \mathrm{Sum}\left(\{\mathrm{COD}_{\mathcal{S}^k}(x_i) : \mathcal{S}^k \in \mathrm{SS}\}\right),$$

- $\mathrm{COD}_{\mathcal{S}^k}(x_i)$: 부분공간 $\mathcal{S}^k$에서 COD를 사용하여 측정한 부분공간 이상정도

위 과정에서 (A)의 COD는 기존의 이상치 탐지기법 중 어떠한 것을 사용해도 무관하며 가장 자주 사용되는 탐지기법은 Breuning 등 (2000)에 소개된 local outlier factor (LOF) 탐지기법이다. LOF는 관측치가 k번째 인접이웃과 떨어진 거리에 기반해 추정한 지역적 밀도를 이용해 이상정도를 정의하는 방법이며 자료에 포함된 변수를 모두 사용하는 대표적인 거리기반 비지도 이상치 탐지 기법이다.

2.1.2. Isolation forest (IF)

IF (Liu, 2008)는 임의의 변수와 임의의 임계값(threshold)을 사용하여 모든 관측치가 끝마디(terminal node)에 고립되는 의사결정 나무를 만들어 부분공간 이상정도를 정의한다. **IF**에서 관측치 x_i의 이상정도를 계산하는 절차는 다음과 같다.

(A) 사용자 정의 요소를 결정한다.

- m: 의사결정 나무의 개수

(B) 모든 관측치를 고립시키는 의사결정 나무를 생성한다.

(a) 비복원 붓스트랩 자료(bootstrapped sample without replacement) X^k를 생성한다.

(b) 다음 규칙을 적용하여 X^k로부터 의사결정 나무 T^k를 구성한다.

* 분리규칙: 임의의 변수 X_j와 임의의 임계값 $c \in [\min(X_j), \max(X_j)]$를 사용한다.
* 정지규칙: 모든 끝마디에 자료의 개수가 1이하이면 정지한다.

(C) 관측치의 이상정도를 계산한다.

$$\mathrm{IF}(x_i) = 1 - \mathrm{Average}\left(\{\mathrm{PathLength}_{T^k}(x_i) : k \le m\}\right),$$

- $\mathrm{PathLength}_{T^k}(x_i)$: T^k에서 x_i가 속한 끝마디의 깊이(depth)

위 과정에서 (C)에 표기된 끝마디의 깊이는 관측치가 고립될 때까지 사용된 분리(split)의 수를 의미한다. **IF**는 정상치는 밀집되어 있어 고립에 필요한 분리 개수가 많은 반면 이상치는 대부분의 관측치로부터 떨어져 있어 고립에 필요한 분리 개수가 상대적으로 적다는 가정에 기반한다.

2.2. 지역적 부분공간 탐색(Local subspace search)

지역적 부분공간 탐색은 관측치마다 이상정도 측정에 적합한 부분공간(relevant subspace)을 선정하여 부분공간 이상정도를 측정하는 기법이다. 선정된 부분공간이 여러 개일 경우에는 적절한 결합함수를 이용해 최종 이상정도를 정의한다.

2.2.1. Subspace outlier degree (SOD)

SOD (Kriegel 등, 2009)는 관측치마다 한 개의 적합한 부분공간을 선정하여 관측치의 부분공간 이상정도를 측정한다. **SOD**에서 관측치 x_i의 이상정도를 계산하는 절차는 다음과 같다.

(A) 사용자 정의 요소를 결정한다.

- k: 인접이웃의 개수
- l: 공유된 인접이웃의 개수
- α: 거리제곱의 비율

(B) 공유된 인접이웃(shared nearest neighbor) 집합을 결정한다.

$$\mathrm{SN}(x_i) = \{x_t : S_k(x_i, x_t) \ge \mathrm{Upper}_\ell(\{S_k(x_i, x_s) : s \ne i\})\},$$

- $S_k(x_i, x_t) = \left|N_k(x_i) \cap N_k(x_t)\right|$
- $N_k(x_i) = \{x_t : \mathrm{Dist}_{\mathcal{D}}(x_i, x_t) \le \mathrm{Lower}_k(\{\mathrm{Dist}_{\mathcal{D}}(x_i, x_s) : s \ne i\})\}$

(C) x_i의 적합한 부분공간을 선정한다.

$$\mathcal{S}_\alpha(x_i) = \left\{ j : \sigma^2_{\{j\}}(x_i) < \frac{\alpha \sigma^2_{\mathcal{D}}(x_i)}{d},\ j \in \mathcal{D} \right\},$$

- $\sigma^2_{\{j\}}(x_i) = \mathrm{Average}\left(\left\{\mathrm{Dist}_{\{j\}}(x_t, \mu(x_i))^2 : x_t \in \mathrm{SN}(x_i)\right\}\right)$
- $\sigma^2_{\mathcal{D}}(x_i) = \mathrm{Average}\left(\left\{\mathrm{Dist}_{\mathcal{D}}(x_t, \mu(x_i))^2 : x_t \in \mathrm{SN}(x_i)\right\}\right)$
- $\mu(x_i) = \mathrm{Average}\left(\{x_t : x_t \in \mathrm{SN}(x_i)\}\right)$

(D) 관측치의 이상정도를 계산한다.

$$\mathrm{SOD}(x_i) = |\mathcal{S}_\alpha(x_i)|^{-1}\mathrm{Dist}_{\mathcal{S}_\alpha(x_i)}(x_i, \mu(x_i)).$$

위 과정에서 (B)의 $\mathrm{SN}(x_i)$는 x_i의 k개의 인접이웃 집합 $N_k(x_i)$와 다른 관측치들의 k개의 인접이웃 집합에 대하여 공유된 인접이웃 유사도(shared nearest neighbor similarity) S_k로 결정한 l개의 공유된 인접이웃 집합이다. 공유된 인접이웃 유사도는 유사도 측정에 관측치 사이의 거리를 직접 사용하지 않고 거리의 순위 정보를 이용하기 때문에 고차원 자료에서 집중효과가 존재할 때 유사도를 안정적으로 측정할 수 있는 장점을 가진다 (Houle 등, 2010). (C)에서 $\mathcal{S}_\alpha(x_i)$는 x_i의 이상정도 측정에 적합한 부분공간으로, 공유된 인접이웃 집합 $\mathrm{SN}(x_i)$에 포함된 관측치의 중심 $\mu(x_i)$와 각 관측치 사이의 일차원 변수공간 거리의 제곱의 평균 $\sigma^2_{\{j\}}$이 전체공간 거리의 제곱의 평균 $\sigma^2_{\mathcal{D}}$의 α배보다 작은 변수들의 모임으로 구성된다. 이러한 부분공간 선정 원리는 공유된 인접이웃을 일차원 변수공간에 투사(projection)했을 때 대다수의 관측치들이 밀집해 있는 경우 소수의 관측치가 떨어진 정도를 잘 관측할 수 있다는 관찰에 기인한다. 최종 이상정도는(D)와 같이 정의되며 이는 부분공간 $\mathcal{S}_\alpha(x_i)$에서 x_i가 $\mu(x_i)$로부터 떨어진 거리를 의미한다. **SOD**는 이상치 탐지에 적합한 부분공간을 관측치마다 고유하게 선정한다는 점에 의의가 있다. 하지만 부분공간 선정 단계에서 일차원 변수공간의 정보만을 이용하기 때문에 변수간의 의존도(dependency)를 고려하지 못하는 단점이 있다. 또한 한 개의 부분공간만을 선정하기 때문에 복수의 부분공간에서 높은 이상정도를 갖는 관측치를 파악하지 못하는 한계가 있다.

2.2.2. Outlier ranking in relevant subspaces (OUTRES)

OUTRES (Müller 등, 2011)는 관측치마다 이상치 탐지에 적합한 부분공간을 여러개 선정해 부분공간 이상정도를 측정하고 이를 결합하여 최종 이상정도를 정의한다. 어떤 부분공간 $\mathcal{S}$를 관측치 x_i의 적합한 부분공간으로 판단하는 기준은 다음의 인접이웃 집합에 기반해 정의된다.

$$N_{\mathcal{S},\varepsilon}(x_i) = \{x_t : \mathrm{Dist}_{\mathcal{S}}(x_i, x_t) \le v_\varepsilon(|\mathcal{S}|),\ t \ne i\}, \quad \mathcal{S} \subset \mathcal{D}, \varepsilon > 0,$$

단, $v_\varepsilon(z) = \epsilon h(z)/h(2)I(z > 2) + \varepsilon I(z \le 2)$이고 $h(z) = \{8(z+4)(2\sqrt{\pi})^z\pi^{-z/2}\Gamma(z/2+1)\}n^{-1/z+4}$이다. ε은 인접이웃의 범위를 결정하는 상수이며 $h(z)$는 z에 대해 단조증가하는 함수로 차원이 커질수록 관측치들의 평균거리가 길어져 인접이웃 집합이 듬성듬성(sparse)해지는 문제를 조정하기 위해 사용되었다 (Silverman, 1986). **OUTRES**는 인접이웃 집합 $N_{\mathcal{S},\varepsilon}$에 포함된 관측치가 균등하게 분포되어 있는지에 대하여 유의수준을 $\alpha \in (0,1)$로 하는 적합도 검정(goodness of fit test)을 시행하며, 귀무가설이 기각되는 경우 $\mathcal{S}$를 x_i의 적합한 부분공간으로 판단한다. 검정 방법으로는 콜모고로프 스미노르프 적합도 검정(Kolmogorov-Smirnov goodness-of-fit test)이 사용된다 (Stephens, 1970). 관측치 x_i에 대한 적합한 부분공간이 결정되면 이 부분공간에서 커널 밀도함수 추정기법(kernel density estimation)을 사용하여 x_i 의 정상정도(degree of normality)를 측정한다. 이때 사용되는 커널 함수는 $K(x) = (1-x^2)I(x < 1)$이다. **OUTRES**에서 관측치 x_i의 이상정도를 측정하는 절차는 다음과 같다. 주어진 부분공간 $\mathcal{S}$에 대하여 위 검정의 귀무가설이 기각되면, 즉 $\mathcal{S}$가 적합한 부분공간으로 판단되면 $\mathrm{KS}_{\mathcal{S},\alpha}(x_i) = 1$이라고 하자.

(A) 사용자 정의 요소를 설정한다.

- ε: 인접이웃 범위 설정 상수
- $\alpha \in (0,1)$: 검정의 유의수준

(B) x_i의 적합한 부분공간 집합을 탐색한다.

$$\mathrm{SS}(x_i) = \left\{\mathrm{SS}^k(x_i) : k \le d\right\},$$

- $\mathrm{SS}^1(x_i) = \{\mathcal{S} : \mathrm{KS}_{\mathcal{S},\alpha}(x_i) = 1, |\mathcal{S}| = 1\}$
- $\mathrm{SS}^k(x_i) = \left\{\mathcal{S} \cup \mathcal{S}' : \mathrm{KS}_{\mathcal{S}\cup\mathcal{S}',\alpha}(x_i) = 1, |\mathcal{S} \cup \mathcal{S}'| = k,\ \mathcal{S}, \mathcal{S}' \in \mathrm{SS}^{k-1}(x_i)\right\},\ 2 \le k \le d$

(C) 부분공간 $\mathcal{S} \in \mathrm{SS}(x_i)$에서 관측치의 정상정도를 정의한다.

$$\mathrm{Score}_{\mathcal{S}}(x_i) = \begin{cases} \dfrac{\mathrm{Den}_{\mathcal{S}}(x_i)}{\mathrm{Dev}_{\mathcal{S}}(x_i)}, & \text{if } \mathrm{Dev}_{\mathcal{S}}(x_i) \ge 1, \\ 1, & \text{else.} \end{cases}$$

- $\mathrm{Den}_{\mathcal{S}}(x_i) = \mathrm{Average}\left(\left\{K\left(\mathrm{Dist}_{\mathcal{S}}(x_i, x_t)\right)/v_\varepsilon(|\mathcal{S}|) : x_t \in N_{\mathcal{S},\varepsilon}(x_i)\right\}\right)$
- $\mathrm{Dev}_{\mathcal{S}}(x_i) = \left(\mu_{\mathcal{S}}(x_i) - \mathrm{Den}_{\mathcal{S}}(x_i)\right)/2\sigma_{\mathcal{S}}(x_i)$
  - $\sigma_{\mathcal{S}}(x_i) = \mathrm{Average}\left(\left\{(\mathrm{Den}_{\mathcal{S}}(x_t) - \mu_{\mathcal{S}}(x_i))^2 : x_t \in N_{\mathcal{S},\varepsilon}(x_i)\right\}\right)^{1/2}$
  - $\mu_{\mathcal{S}}(x_i) = \mathrm{Average}\left(\{\mathrm{Den}_{\mathcal{S}}(x_t) : x_t \in N_{\mathcal{S},\varepsilon}(x_i)\}\right)$

(D) 최종 이상정도를 계산한다.

$$\mathrm{OUTRES}(x_i) = 1 - \mathrm{Product}(\{\mathrm{Score}_{\mathcal{S}}(x_i) : \mathcal{S} \in \mathrm{SS}(x_i)\}).$$

위 과정에서 (B)는 계산비용을 줄이기 위해 도입된 연관규칙(apriori) 알고리즘 (Agrawal와 Srikan, 1994) 기반의 상향식 부분공간 탐색 과정이다. 이러한 알고리즘이 사용되는 이유는 차원이 커짐에 따라 관측치들이 흩어지는 경향이 나타나기 때문이다 (Beyer 등, 1999). 따라서 만약 어떤 부분공간이 균등하게 분포되었다고 판단되면 더이상의 부분공간 탐색에 해당 부분공간을 고려하지 않는다. (C)는 부분공간 $\mathcal{S}$에서 x_i의 부분공간 정상정도 $\mathrm{Score}_{\mathcal{S}}(x_i)$를 정의한다. $\mathrm{Score}_{\mathcal{S}}(x_i)$는 인접이웃 집합 $N_{\mathcal{S},\varepsilon}(x_i)$에 대해 커널 밀도함수 추정기법으로 추정한 지역적 밀도(local density) $\mathrm{Den}_{\mathcal{S}}(x_i)$와 지역적 밀도의 편차 $\mathrm{Dev}_{\mathcal{S}}(x_i)$의 비율로 계산된다. 커널 밀도 추정에 단조 증가하는 대역폭(bandwidth) $v_\varepsilon(|\mathcal{S}|)$을 사용함으로써 차원크기가 다른 부분공간에서 추정된 밀도의 스케일을 균일하게 한다. **OUTRES**는 관측치마다 여러 개의 적합한 부분공간을 탐색할 수 있는 알고리즘을 제안하고 차원 크기가 다른 부분공간에서 측정된 이상정도를 통합할 때 발생할 수 있는 스케일의 문제를 보완했다는 점에 의의가 있다.

2.2.3. Outlier ranking via subspace analysis (OutRank)

OutRank (Müller 등, 2012)는 부분공간 클러스터링(subspace clustering)기법을 이용해 관측치마다 적합한 부분공간을 선정하여 이상정도를 측정하는 기법이다. 부분공간 클러스터링은 상향식 혹은 하향식 알고리즘에 기반해 자료의 부분공간에 내재되어 있는 클러스터를 탐색하는 기법 (Parsons 등, 2004)이다. **OutRank**에서 관측치 x_i의 적합한 부분공간을 선정하고 이상정도를 측정하는 절차는 다음과 같다.

(A) 사용자 정의 요소를 결정한다.

- SC: 부분공간 클러스터링 기법

(B) SC를 사용하여 자료의 부분공간 클러스터 집합을 생성한다.

$$\mathrm{SCR} = \{(C_k, \mathcal{S}_k) : k \le m\},$$

- m: 부분공간의 개수
- C_k: 부분공간 $\mathcal{S}_k$에서 구성한 클러스터 집합

(C) x_i의 부분공간 정상정도를 측정한다.

$$\text{Regular}(x_i) = \text{Average}\left(\{\text{Evid}_k(x_i) : k \in \text{SCR}(x_i)\}\right),$$

- $\text{SCR}(x_i) = \{k : |C_k(x_i)| \neq 0, k \leq m\}$
- $C_k(x_i)$: C_k에 포함된 클러스터 중 x_i가 속한 클러스터
- $\text{Evid}_k(x_i) = |C_k(x_i)|/\max_{j\leq m}|C_j| + |\mathcal{S}_k|/\max_{j\leq m}|\mathcal{S}_j|$

(D) 최종 이상정도를 계산한다.

$$\text{OutRank}(x_i) = 1 - \text{Regular}(x_i).$$

(A)에서 SC는 어떠한 부분공간 클러스터링 기법을 사용해도 무관하며 Procopiuc 등 (2002)가 개발한 density-based optimal projective clustering (DOC)가 가장 대표적인 방법 중 하나이다. (B)의 SCR은 자료에 SC를 적용하여 얻은 부분공간 클러스터 집합이다. (C)에서 $\text{Evid}_k(x_i)$는 부분공간 $\mathcal{S}_k$에서의 클러스터 집합 C_k에 x_i가 속해있는 정도(degree of cluster membership)를 나타낸다. SCR에 속한 부분공간 클러스터 집합 중 x_i를 포함하고 있는 클러스터에서의 $\text{Evid}_k(x_i)$값을 더하여 x_i의 정상정도 $\text{Regular}(x_i)$를 정의한다. 이 정상정도는 x_i가 다수의 부분공간 클러스터에 포함되어 있고 해당 클러스터의 크기와 부분공간의 차원이 클수록 큰 값을 가진다. **OutRank**는 서로 다른 클러스터가 많은 수의 공통된 관측치를 공유(overlap)하는 경우 정상정도를 과대계상 하는 현상이 나타남을 지적하며 이를 보완할 수 있는 Evid_k 함수를 추가로 제안하였다. 이에 대한 구체적인 설명은 Müller 등 (2012)에 기재되어 있다. **OutRank**는 부분공간 클러스터링 알고리즘을 부분공간 이상치 탐지에 직접 적용할 수 있는 방법을 제안했다는 점에 의의가 있다. 또한 다양한 부분공간 클러스터링 기법을 선택할 수 있어 방법론 적용에 유연성을 가진다.

2.3. 전역적 부분공간 탐색(Global subspace search)

전역적 부분공간 탐색기법은 자료의 모든 관측치에 공통되게 적용되는 적합한 부분공간을 탐색하며 최종 이상정도는 탐색된 모든 부분공간에서의 이상정도를 결합하여 계산된다. 이 기법은 부분공간 선정 단계와 이상정도 측정 단계가 분리되어 이상정도 측정에 기존의 이상치 탐지기법을 차용할 수 있기 때문에 이상치 탐지를 위한 전처리 과정으로 생각할 수도 있다.

2.3.1. High contrast subspaces (HiCS)

HiCS (Keller 등, 2012)는 이차원 이상의 부분공간에서 변수 사이의 상호 상관관계(mutual correlation)의 정도를 측정하는 Contrast의 개념을 소개하고 이 값이 큰 고 대비 부분공간(high contrast subspace)을 이상치 탐지에 적합한 부분공간으로 선정한다. **HiCS**는 일차원 변수공간에서는 밀집되어 있어 이상치로 보이지 않지만 이차원 이상의 변수공간에서 이상정도를 드러내는 자명하지 않은 이상치(non-trivial outlier)를 정의하고 고 대비 부분공간에서 이러한 이상치를 탐지하기 유용함을 보인다.

Contrast는 특정 부분공간에 속한 변수의 주변 분포함수와 해당 변수를 제외한 변수들을 특정 영역으로 제한했을 때의 조건부 분포함수와의 차이에 기반해 정의된다. 이는 다른 변수들과 상호독립(mutually independent)인 변수라면 두 값의 차이가 없을 것이라는 가정에 기반한다. **HiCS**의 이상치 탐지 절차를 소개하기 전에 Contrast를 부분공간에 대한 함수로 정의하면 다음과 같다.

$$\text{Contrast}(\mathcal{S}) = \text{Average}\left(\left\{\text{Dev}(X_j, X_j|R_j) : j \in \mathcal{S}'\right\}\right), \quad \mathcal{S} \subset \mathcal{D},\ |\mathcal{S}| \geq 2.$$

- $\mathcal{S}' = \{j_1, \ldots, j_m\}$: $\mathcal{S}$에서 복원 추출한 m개의 변수 인덱스 집합.

- $R_j = \prod_{l \in \mathcal{S}, l \neq j} [a_l, b_l]$: 다음을 만족하는 임의의 구간 $[a_l, b_l]$으로 구성한 격자 부분집합

$$\text{Average}\left(\{I(x_{il} \in [a_l, b_l]) : i \leq n\}\right) \geq \alpha^{-\frac{1}{|\mathcal{S}|}}, \quad \alpha \in (0, 1)$$

- $\text{Dev}(X_j, X_j|R_j) = \sup_{x_{ij} \in X_j|R_j} \left|F_{X_j}(x_{ij}) - F_{X_j|R_j}(x_{ij})\right|$
  - $X_j|R_j = \left\{x_{ij} : x_{il} \in [a_l, b_l],\ l \in \mathcal{S}, l \neq j\right\}$
  - $F_{X_j}(x_{ij}) = \text{Average}\left(\left\{I(x_{sj} < x_{ij}) : x_{sj} \in X_j\right\}\right)$
  - $F_{X_j|R_j}(x_{ij}) = \text{Average}\left(\left\{I(x_{sj} < x_{ij}) : x_{sj} \in X_j|R_j\right\}\right)$

위 정의에서 $F_{X_j}(x_{ij})$와 $F_{X_j|R_j}(x_{ij})$는 X_j와 $X_j|R_j$에 대한 경험적 누적분포(empirical cumulative distribution)함수의 x_{ij}에서의 함수값이다. **HiCS**의 적합한 부분공간 탐색과 이상정도 측정 절차는 다음과 같다.

(A) 사용자 정의 요소를 결정한다.

- COD: 부분공간에서 사용할 이상치 탐지기법
- $\alpha \in (0, 1)$: Contrast함수에서 구간 $[a_l, b_l]$에 포함되는 자료의 비율
- m: Contrast측정에 사용하는 Dev함수의 개수
- t: 각 부분공간 탐색 단계에서 선정하는 부분공간의 개수

(B) 자료의 적합한 부분공간 집합을 탐색한다.

$$\text{SS} = \{\text{SS}^k : 2 \leq k \leq d\},$$

- $\text{SS}^2 = \{\mathcal{S} : \text{Contrast}(\mathcal{S}) \geq \text{Upper}_t(\{\text{Contrast}(\mathcal{S}) : |\mathcal{S}| = 2\}), |\mathcal{S}| = 2\}$
- $\text{SS}^k = \left\{\mathcal{S} \cup \mathcal{S}' : \text{Contrast}(\mathcal{S} \cup \mathcal{S}') \geq \text{Upper}_t\left(\left\{\text{Contrast}(\mathcal{S} \cup \mathcal{S}') : |\mathcal{S} \cup \mathcal{S}'| = k,\ \mathcal{S}, \mathcal{S}' \in \text{SS}^{k-1}\right\}\right), |\mathcal{S} \cup \mathcal{S}'| = k\right\}$, $3 \leq k \leq d$
- $\mathcal{S}, \mathcal{S}' \in \text{SS}$에 대해, $\mathcal{S}' \subset \mathcal{S}$ 그리고 $\text{Contrast}(\mathcal{S}') < \text{Contrast}(\mathcal{S})$에 해당하는 $\mathcal{S}'$을 제거한다.

(C) 관측치의 최종 이상정도를 다음과 같이 계산한다.

$$\text{HiCS}(x_i) = \text{Average}\left(\{\text{COD}_{\mathcal{S}}(x_i) : \mathcal{S} \in \text{SS}\}\right),$$

- $\text{COD}_{\mathcal{S}}(x_i)$: 부분공간 $\mathcal{S}$에서 COD를 사용하여 계산한 부분공간 이상정도

위 과정에서 (A)의 COD는 기존의 이상치 탐지기법 중 어떠한 것을 사용해도 무관하다. (B)는 계산비용을 줄이기 위해 제안된 빔 탐색(beam search) (Steinbiss 등, 1994) 기반 상향식 부분공간 탐색 알고리즘이다. 이러한 알고리즘이 사용되는 이유는 어떤 부분공간이 클러스터 구조를 가질 경우 해당 부분공간의 저차원 부분공간에서도 클러스터를 이루는 경향이 나타나기 때문이다 (Agrawal 등, 1998). 최종 이상정도는 (C)와 같이 정의된다. **HiCS**는 이차원 이상의 변수공간에서의 상호 상관관계를 측정하는 새로운 측도 Contrast의 개념을 제안했다는 점에 의의가 있다.

2.3.2. Cumulative mutual information (CMI)

CMI (Nguyen 등, 2013)는 **HiCS**의 발전된 기법으로, 부분공간에 포함된 변수 사이의 상호 상관관계를 측정하는 새로운 측도 CMI를 제안한다. CMI는 부분공간 탐색 과정에서 Contrast와 달리 계산의 임의성을 배제하며 누적분포함수 대신 누적엔트로피에 기반해 계산된다. **CMI**의 이상치 탐지 절차를 소개하기 전에 CMI를 부분공간에 대한 함수로 정의하면 다음과 같다.

$$\mathrm{CMI}(\mathcal{S}) = \mathrm{Average}\left(\left\{\mathrm{Diff}\left(X_{j_k}, X_{j_k}|\mathrm{Clust}\left(X_{\mathcal{S}_{k-1}}\right)\right) : 2 \le k \le |\mathcal{S}|\right\}\right), \quad \mathcal{S} \subset \mathcal{D},$$

- $\mathcal{S}_1 = \{j_1\}, \mathcal{S}_2 = \{j_1, j_2\}$
  - $\{j_1, j_2\} = \arg\max_{s \ne t \in \mathcal{S}} \mathrm{Diff}\,(X_s, X_s|\mathrm{Clust}(X_t))$
    - $\mathrm{Clust}(X_t) = \{C_{t1}, \ldots, C_{tm_t}\}$: X_t로 구성한 m_t개의 클러스터 집합
  - $\mathrm{Diff}(X_s, X_s|\mathrm{Clust}(X_t)) = \mathrm{CumEnt}(X_s) - \sum_{l=1}^{m_t}(|C_{tl}|/n)\mathrm{CumEnt}(X_s|C_{tl})$
    - $\mathrm{CumEnt}(X_s) = -\sum_{i \le |X_s|-1}(x_{(i+1)s} - x_{(i)s})(i/|X_s|)\log(i/|X_s|)$
      - $x_{(i)s} = \mathrm{Upper}_i\,(X_s)$
    - $\mathrm{CumEnt}(X_s|C_{kl}) = -\sum_{i \le |X_s|C_{kl}|-1}(x_{(i+1)s} - x_{(i)s})(i/|X_s|C_{kl}|)\log(i/|X_s|C_{kl}|)$
      - $X_s|C_{tl} = \{x_{is} : i \in C_{tl}\}$
- $\mathcal{S}_{k+1} = \{j_s : s \le k+1\}, 2 \le k \le |\mathcal{S}| - 1$
  - $j_{k+1} = \arg\max_{s \in \mathcal{S}, s \notin \mathcal{S}_k} \mathrm{Diff}(X_s, X_s|\mathrm{Clust}(X_{\mathcal{S}_k}))$
    - $\mathrm{Clust}(X_{\mathcal{S}_k}) = \{C_{k1}, \ldots, C_{km_k}\}$: $X_{\mathcal{S}_k}$로 구성한 m_k개의 클러스터 집합
  - $\mathrm{Diff}(X_s, X_s|\mathrm{Clust}(X_t, t \in \mathcal{S}_k)) = \mathrm{CumEnt}(X_s) - \sum_{l=1}^{m_k}(|C_{kl}|/n)\mathrm{CumEnt}(X_s|C_{kl})$
    - $X_s|C_{kl} = \{x_{is} : i \in C_{kl}\}$

위 정의에서 $\mathrm{CumEnt}(X_s)$와 $\mathrm{CumEnt}(X_s|C_{kl})$은 변수 X_s와 제한된 변수 $X_s|C_{kl}$에 대한 경험적 누적엔트로피(empirical cumulative entropy) 함수이다. CMI는 임의성에 의존하지 않지만 계산에 사용되는 변수의 선택 순서에 따라 값이 달라지므로 이를 보완하기 위해 주어진 부분공간 $\mathcal{S}$에서 근사적으로 최대의 CMI값을 산출한다.

CMI에서 자료의 적합한 부분공간을 탐색하고 이상정도를 측정하는 절차는 다음과 같다.

(A) 사용자 정의 요소를 결정한다.

- COD: 부분공간에서 사용할 이상치 탐지기법
- Clust: 부분공간에서 사용할 클리스터링 기법
- t: 각 부분공간 탐색 단계에서 선정하는 부분공간의 개수

(B) 자료의 적합한 부분공간 집합을 선정한다.

$$\mathrm{SS} = \{\mathrm{SS}^k : 2 \le k \le d\},$$

- $\mathrm{SS}^2 = \{\mathcal{S} : \mathrm{CMI}(\mathcal{S}) \ge \mathrm{Upper}_t(\{\mathrm{CMI}(\mathcal{S}) : |\mathcal{S}| = 2\}), |\mathcal{S}| = 2\}$
- $\mathrm{SS}^k = \{\mathcal{S} \cup \mathcal{S}' : \mathrm{CMI}(\mathcal{S} \cup \mathcal{S}') \ge \mathrm{Upper}_t(\{\mathrm{CMI}(\mathcal{S} \cup \mathcal{S}') : |\mathcal{S} \cup \mathcal{S}'| = k,\ \mathcal{S}, \mathcal{S}' \in \mathrm{SS}^{k-1}\}), |\mathcal{S} \cup \mathcal{S}'| = k\}, 3 \le k \le d$
- $\mathcal{S}, \mathcal{S}' \in \mathrm{SS}$에 대해, $\mathcal{S}' \subseteq \mathcal{S}$ 그리고 $\mathrm{CMI}(\mathcal{S}') < \mathrm{CMI}(\mathcal{S})$에 해당하는 $\mathcal{S}'$을 제거한다.

Table 1: Summary of subspace outlier detection methods

Method	Subspace search		Subspace result		Outlierness function		Implementation tool
	Criteria	Algorithm	Local vs Global	Single vs Multiple	Subspace outlierness	Combination	
FB	Random	None	Global	Multiple	$\mathrm{COD}_{\mathcal{S}^k}(x_i)$	Sum	ELKI
IF	Random	None	Global	Multiple	$\mathrm{PathLength}_{T^k}(x_i)$	1-Average	R
SOD	Distance from mean on one diemensional space	Brute force	Local	Single	$\mathrm{SOD}(x_i)$	None	ELKI
OUTRES	Goodness-of-fit test for uniform distribution	Apriori	Local	Multiple	$\mathrm{Score}_{\mathcal{S}}(x_i)$	1 − Product	ELKI
OutRank	Depends on the subspace clustering	Depends on the subspace clustering	Local	Multiple	$\mathrm{Evid}_k(x_i)$	1 − Average	ELKI
HiCS	Contrast	Beam search	Global	Multiple	$\mathrm{COD}_{\mathcal{S}}(x_i)$	Average	ELKI
CMI	CMI	Beam search	Global	Multiple	$\mathrm{COD}_{\mathcal{S}}(x_i)$	Average	Java

(C) 관측치의 최종 이상정도를 다음과 같이 계산한다.

$$\mathrm{CMI}(x_i) = \mathrm{Average}\left(\{\mathrm{COD}_{\mathcal{S}}(x_i) : \mathcal{S} \in \mathrm{SS}\}\right),$$

- $\mathrm{COD}_{\mathcal{S}}(x_i)$: 부분공간 $\mathcal{S}$에서 COD를 사용하여 계산한 부분공간 이상정도

위 과정에서 (A)의 COD는 기존의 이상치 탐지기법 중 어떠한 것을 사용해도 무관하다. (B)는 자료의 적합한 부분공간을 근사적으로 탐색하는 알고리즘으로 **HiCS**와 동일하게 빔 탐색(beam search)에 기반한 상향식 부분공간 탐색 알고리즘이 사용된다. 최종 이상정도는 (C)와 같이 정의 된다.

2.4. 부분공간 이상치 탐지기법의 구현 도구

Table 1은 앞서 살펴본 부분공간 이상치 탐지기법들을 부분공간의 선정 방식, 부분공간 탐색의 결과, 이상정도 측정함수 그리고 이를 적용할 수 있는 컴퓨팅 프로그램을 기준으로 요약하여 나타낸 표이다. 대다수의 방법론들은 오픈 소스 데이터 마이닝 소프트웨어인 environment for developing kdd-applications supported by index-structures (ELKI) (Schubert와 Zimek, 2019)를 이용해 적용할 수 있다. ELKI는 클러스터링과 이상치 탐지를 주 목적으로 개발된 Java기반의 소프트웨어로 R*-tree (Beckmann, 1990)와 같은 인덱스 구조(index structure)에 기반해 인접이웃 계산 등을 효율적으로 수행하는 장점이 있다. ELKI는 소스 코드 뿐만 아니라 사용자가 손쉽게 이용할 수 있도록 graphical user interface (GUI) 소프트웨어를 배포하기 때문에 높은 접근성을 가지는 장점이 있다. **IF**는 R의 h2o 패키지에 포함된 h2o.isolationForest함수를 이용해 적용할 수 있으며 **CMI**는 CMI의 웹사이트 에서 제공하는 Java 기반의 소스코드를 사용하여 적용할 수 있다.

2.5. 부분공간 이상치 탐지기법의 활용과 평가

비지도 이상치 탐지기법은 관측치의 이상치 여부(outlier label)가 자료에 포함되어 있지 않은 경우에 사용된다. 따라서 이상치를 직접 결정하는 방법론이기 보다는 이상정도에 기반해 관측치의 순위를 제공하여 자료 전문가의 경험적 판단을 보조하는 수단으로 사용되는 경우가 많다. 이상치 여부를 결정할 수 있도록 도출된 이상정도에 대하여 일반적인 임계치(threshold)를 설정할 수도 있으나 이는 자료의 특징과 무관한 이상치 탐지 결과로 이어질 수 있기 때문에 매우 신중하게 결정되어야 한다. 또한 이상치를 결정하는 정량적인 기준이 없으므로 부분공간 탐색횟수, 거리 측도의 선택, 인접 관측치의 개수 등 각 방법론에서 사용자가 정의해야 하는 요소들을 결정하는 방식도 매우 주관적이다.

하지만 연구 논문 등에서 새로운 방법론을 소개하는 경우 방법론의 우수성을 입증할 필요가 있으므로 이상치 여부가 포함된 자료를 사용하여 receive operating characteristic (ROC)곡선의 area under the curve (AUC)

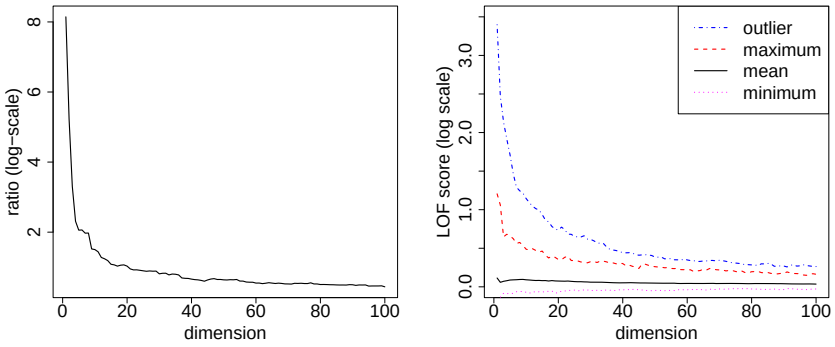

Figure 2: *Two trajectories from the simulations: distance ratio (left) and outlierness measured by LOF with $k = 5$ (right) for the outlier, minimum, maximum and mean of the other observations.*

와 같은 이진 분류(binary classification) 평가 측도를 사용하여 평가하는 경우가 많다. 특히 이상치 탐지에 사용되는 자료는 이상치가 매우 적게 관측되는 불균형 자료(unbalanced sample)인 경우가 대부분이기 때문에 F_1점수 (Powers, 2020) 와 같이 불균형 자료의 분류 결과를 평가하는 측도가 차용되는 경우가 많다. 분류 평가 측도를 사용하는 경우 사용자 결정 요소를 다양하게 적용할 수 있고 이 중 분류의 결과가 최적이 되는 요소를 결정하는 것이 가능하며 이러한 방식을 통하여 사용자 결정 요소의 범위나 값을 논문에 함께 추천하기도 한다.

3. 가상실험 및 자료 분석

3.1. 집중효과에 대한 가상실험

집중효과가 발생할 경우 이상정도가 어떻게 왜곡되는지 확인하기 위한 가상실험을 진행하였다. 가상실험에 사용된 자료는 $N(0, 1)$을 따르는 서로 독립인 d개의 확률변수에 대하여 $n = 1000$개의 임의 표본을 생성하여 구성하였다.

먼저 이 자료에서 원점과 관측치 사이의 최대 거리와 최소 거리의 비율 R이 차원의 크기 d에 따라 어떻게 변화하는지 조사하였으며 그 결과를 세로 축을 $\log R$, 가로 축을 d로 하여 Figure 2의 왼쪽 그래프에 도시하였다. 그림에서 d 즉, 차원이 커짐에 따라 $\log R$의 값이 0으로 수렴하는 것을 확인할 수 있다.

다음으로 자료에 이상치를 포함시켜 관측치의 이상정도가 어떻게 변화하는지 확인하였다. 앞에서 생성한 자료에서 $x_{11} = 10$으로 두어 첫 번째 관측치 x_1이 첫 번째 변수로 인하여 이상치가 되는 경우를 고려하였으며 LOF를 적용하여 각 관측치 x_i의 이상정도 O_i, $i \le n$를 측정하였다. Figure 2의 오른쪽 그래프는 차원의 크기 d에 따라 $\log O_1$, $\min_{i\neq 1} \log O_i$, $\max_{i\neq 1} \log O_i$, $\text{Average}(\{\log O_i : i \neq 1\})$의 값이 어떻게 변화하는지 나타낸다. 그림에서 d의 크기가 작을 때는 x_1의 이상치의 이상정도가 다른 관측치의 이상정도보다 매우 크지만 d의 크기가 커지면서 집중효과가 발생하면 x_1과 다른 관측치의 이상정도가 비슷해지는 것을 확인할 수 있다.

3.2. 자료 분석

본 논문에서 소개한 이상치 탐지기법을 실제 자료에 적합하여 그 결과를 요약하였다. 분석에 사용된 자료는 Campos 등 (2016)에 소개된 이상치 탐지기법 평가용 자료이며 이상치 여부가 포함되어 있는 공개된 자료이다. Table 2는 분석에 사용한 자료를 요약하여 나타낸다.

각 자료에 대하여 논문에서 소개한 부분공간 이상치 탐지기법을 적용하였으며 Tukey (1977)에서 소개한 방법을 이용해 이상치 결정 기준을 설정하였다. 참고로 Tukey (1977)는 관측치의 값이 다음의 범위를 벗어나

Table 2: Datasets used in the experiments

Dataset	Dimension	Sample size	Number of outliers	Percentage of outliers
Glass	7	214	9	4.21
Pima	8	526	26	4.94
WBC	9	454	10	2.20
Lymphography	19	148	6	4.05
Ionosphere	32	351	126	35.90
WPBC	33	198	47	23.74
SpamBase	57	2934	146	4.98
Arrhythmia	259	256	12	4.69

Table 3: F_1-score results

Dataset	LOF	FB	IF	SOD	OUTRES	OutRank	HiCS
Glass	**0.2500**	0.1333	0.1429	0.2381	0.1395	0.0000	0.1538
Pima	**0.1800**	0.0741	0.1176	0.0833	0.0000	0.0000	0.0952
WBC	0.3721	0.1739	0.3000	**0.4167**	0.3462	0.0441	0.3721
Lymphography	0.5882	0.2500	**0.6154**	0.5000	none	0.0714	0.2000
Ionosphere	0.2739	0.1870	0.3007	0.3087	none	0.3422	**0.4285**
WPBC	0.0615	0.0656	0.0000	0.1538	none	**0.2078**	0.1667
SpamBase	0.1184	0.0000	**0.3353**	0.1036	none	0.3115	0.1960
Arrhythmia	0.1764	0.2609	0.2400	0.2500	none	0.1099	**0.3200**

면 해당 관측치를 이상치로 분류한다.

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)].$$

이 때, Q_t, $t \leq 3$는 각각 제 1, 2, 3사분위수이며 보통 $k = 1.5$를 사용한다. 이 방법은 분포 가정을 필요로하지 않으며 극단적인 값의 영향을 적게 받아 강건(robust)하다는 장점을 가진다. 본 논문에서는 자료에 각 방법론을 적용하여 얻은 이상정도에 대해 해당 범위를 계산한 후 그 범위의 오른쪽 값보다 큰 이상정도를 갖는 관측치를 이상치로 판단하였다.

Table 3은 이상치 탐지기법을 적용하여 결정한 이상치 여부와 자료에 포함된 실제 이상치 여부를 이용해 계산한 평가측도 F_1점수를 요약한 표이다. 참고로 변수를 선택하지 않는 경우와 비교하기 위하여 LOF 기법을 사용한 결과도 포함시켰다. 또한 **OUTRES**의 경우 계산량이 너무 많아 24간 이내에 계산을 마치지 못한 자료의 점수를 none으로 표기하였으며 **CMI**는 현재 배포된 버전의 소스코드를 원활히 활용하지 못해 분석에 포함시키지 않았다. 분석 결과를 살펴보면 차원이 낮은 두 자료 Glass와 Pima에선 LOF가 가장 높은 성능을 보이지만 나머지 자료에서는 부분공간 이상치 탐지기법이 더 우월한 성능을 보이는 것을 확인할 수 있다. 다만 부분공간 이상치 탐지기법 중 모든 상황에서 절대 우위를 갖는 방법론은 없었으며 이는 방법론과 자료의 특징에 따라 적절한 이상치 탐지기법을 선택해야 한다는 것으로 이해할 수 있다.

마지막으로 독자의 편의를 위하여 본 논문에서 사용한 사용자 정의 요소의 값을 Table 4에 요약해 두었다. **FB**와 **HiCS**에서 이상치 탐지기법 COD는 LOF를 사용하였으며 LOF와 **SOD**에서 k는 표본 크기의 10%로 설정하였고 **SOD**의 l은 k의 70%로 지정했다. **SOD**, **OUTRES**, **HiCS**의 α 그리고 m은 구현 프로그램의 기본값을 사용하였으며 **OUTRES**의 ϵ은 해당 논문의 권장값인 15를 사용하였다. **OutRank**의 부분공간 클러스터링 기법으로는 DOC를 사용하였다. **FB**와 **IF**의 m 그리고 **HiCS**의 t는 구현 프로그램의 기본값 100을 차원 크기에 따라 적당히 조정하였다.

Table 4: User defined environments used in the experiments

Dataset	LOF	FB	IF	SOD			OUTRES		OutRank	HiCS		
	k	m	m	k	l	α	ϵ	α	SC	α	m	t
Glass	21	20	20	21	15	1.1	15	0.1	DOC	0.1	50	32
Pima	53	20	20	53	37	1.1	15	0.1	DOC	0.1	50	32
WBC	45	20	20	45	32	1.1	15	0.1	DOC	0.1	50	32
Lymphography	15	20	20	15	10	1.1	15	0.1	DOC	0.1	50	32
Ionosphere	35	40	40	35	25	1.1	15	0.1	DOC	0.1	50	100
WPBC	20	40	40	20	14	1.1	15	0.1	DOC	0.1	50	100
SpamBase	293	100	100	293	205	1.1	15	0.1	DOC	0.1	50	100
Arrhythmia	26	200	200	26	18	1.1	15	0.1	DOC	0.1	50	200

4. 결론

본 논문에서는 비지도 부분공간 이상치 탐지기법을 부분공간 선정 방식에 따라 임의적 부분공간 탐색, 지역적 부분공간 탐색, 전역적 부분공간 탐색의 세가지 유형으로 나누어 정리하였다. 또한 각 유형에서 대표적인 방법론을 부분공간 선정 기준과 이상정도 측정 방식에 따라 개괄하고 각 방법론을 적용할 수 있는 컴퓨팅 프로그램을 소개하였다. 더하여 부분공간 이상치 탐지기법의 필요성에 대한 이해를 돕기 위해 집중효과에 대한 간단한 가상실험 결과를 제시하였으며 논문에서 소개한 이상치 탐지기법들을 실제 자료에 적용하여 결과를 비교하였다. 본 논문이 비지도 부분공간 탐지기법을 이해하고 활용하는데 큰 도움이 될 것으로 기대한다.

References

Agrawal R and Srikant R (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference Very Large Data Bases, VLDB*, **125**, 487–499.

Agrawal R, Gehrke J, Gunopulos D, and Raghavan P (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, 94–105.

Barnett V and Lewis T (1984). *Outliers in Statistical Data*(2nd ed), Chichester, Wiley.

Beckmann N, Kriegel HP, Schneider R, and Seeger B (1990). The R*-tree: An efficient and robust access method for points and rectangles. In *Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data*, 322–331.

Bennett KP, Fayyad U, and Geiger D (1999). Density-based indexing for approximate nearest-neighbor queries. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 233–243.

Beyer K, Goldstein J, Ramakrishnan R, and Shaft U (1999). When is “nearest neighbor” meaningful?. In *International Conference on Database Theory*, Springer, Berlin, 217–235.

Breunig MM, Kriegel HP, Ng RT, and Sander J (2000). LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93–104.

Campos GO, Zimek A, Sander J, *et al.* (2016). On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study, *Data Mining and Knowledge Discovery*, **30**, 891–927.

Durrant RJ and Kabán A (2009). When is ‘nearest neighbour’ meaningful: A converse theorem and implications, *Journal of Complexity*, **25**, 385–397.

Eskin E, Arnold A, Prerau M, Portnoy L, and Stolfo S (2002). A geometric framework for unsupervised anomaly

detection, *In Applications of Data Mining in Computer Security*, Springer, Boston, 77–101.

Fawcett T and Provost F (1997). Adaptive fraud detection, *Data Mining and Knowledge Discovery*, **1**, 291–316.

Hawkins DM (1980). *Identification of Outliers*, Chapman and Hall, London.

Houle ME, Kriegel HP, Kröger P, Schubert E, and Zimek A. (2010). Can shared-neighbor distances defeat the curse of dimensionality?. In *International Conference on Scientific and Statistical Database Management*, Springer, Berlin, 482–500.

Keller F, Muller E, and Bohm K (2012). HiCS: High contrast subspaces for density-based outlier ranking. In *2012 IEEE 28th International Conference on Data Engineering*, 1037–1048.

Kriegel HP, Kröger P, Schubert E, and Zimek A (2009). Outlier detection in axis-parallel subspaces of high dimensional data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, Berlin, 831–838.

Lazarevic A and Kumar V (2005). Feature bagging for outlier detection. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 157–166.

Liu FT, Ting KM, and Zhou ZH (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, 413–422.

Müller E, Schiffer M, and Seidl T (2011). Statistical selection of relevant subspace projections for outlier ranking. In *2011 IEEE 27th International Conference on Data Engineering*, 434–445.

Müller E, Assent I, Iglesias P, Mülle Y, and Böhm K (2012). Outlier ranking via subspace analysis in multiple views of the data. In *2012 IEEE 12th International Conference on Data Mining*, 529–538.

Nguyen HV, Müller E, Vreeken J, Keller F, and Böhm K (2013). CMI: An information-theoretic contrast measure for enhancing subspace cluster and outlier detection. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, 198–206.

Parsons L, Haque E, and Liu H (2004). Subspace clustering for high dimensional data: a review, *Acm Sigkdd Explorations Newsletter*, **6**, 90–105.

Penny KI and Jolliffe IT (2001). A comparison of multivariate outlier detection methods for clinical laboratory safety data, *Journal of the Royal Statistical Society: Series D (The Statistician)*, **50**, 295–307.

Powers DM (2020). *Evaluation: from Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation*.

Procopiuc CM, Jones M, Agarwal PK, and Murali TM (2002). A Monte Carlo algorithm for fast projective clustering. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, 418–427.

Schubert E and Zimek A (2019). *ELKI: A Large Open-Source Library for Data Analysis-ELKI Release 0.7. 5" Heidelberg*.

Silverman BW (1986). *Density Estimation for Statistics and Data Analysis*, **26**, CRC press.

Steinbiss V, Tran BH, and Ney H (1994). Improvements in beam search. In *Third International Conference on Spoken Language Processing*.

Stephens MA (1970). Use of the kolmogorov–smirnov, cramer–von mises and related statistics without extensive tables, *Journal of the Royal Statistical Society: Series B (Methodological)*, **32**, 115–122.

Tukey JW (1977). *Exploratory Data Analysis*, **2**, 131–160.

Zimek A, Schubert E, and Kriegel HP (2012). A survey on unsupervised outlier detection in high-dimensional numerical data, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **5**, 363–387.

Received February 16, 2021; Revised March 3, 2021; Accepted March 5, 2021

고차원 자료의 비지도 부분공간 이상치 탐지기법에 대한 요약 연구

안재형[a], 권성훈[1,a]

[a]건국대학교 응용통계학과

요 약

고차원 자료에서 이상치를 탐지하기 위해서는 변수를 선별해야 할 필요성이 있다. 이상치 탐지에 적합한 정보가 종종 일부 변수에만 포함되어 있기 때문이다. 많은 수의 부적합한 변수가 자료에 포함될 경우 모든 관측치의 거리가 비슷해지는 집중효과가 발생하고 이로 인해 모든 관측치의 이상정도가 비슷해지는 문제가 발생하게 된다. 부분공간 이상치 탐지기법은 전체 변수 중 이상치 탐지에 적합한 변수들의 집합을 선별하여 관측치의 이상정도를 측정함으로써 이러한 문제를 극복한다. 본 논문은 대표적인 부분공간 이상치 탐지기법을 부분공간 선정 방식에 따라 세가지 유형으로 분류하고 각 유형에 속한 방법론을 부분공간 선정 기준과 이상정도 측정 방식에 따라 요약한다. 더하여, 부분공간 이상치 탐지기법들을 적용할 수 있는 컴퓨팅 프로그램을 소개하고 집중효과에 대한 간단한 가상 실험과 자료 분석 결과를 제시한다.

주요용어: 이상치 탐지, 고차원 자료, 부분공간 이상치 탐지

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NO. NRF-2020R1F1A1A01071036).

[1]교신저자: 서울시 광진구 능동로 120, 건국대학교 응용통계학과. E-mail: shkwon0522@konkuk.ac.kr