수질자료의 이상치 탐색을 위한 Isolation Forest기법의 적용

Application of Isolation Forest Technique for Outlier Detection in Water Quality Data

김종은 · 박노석 · 윤상진 · 채선하* · 윤석민[†]

Jongeun Kim · No-Suk Park · Sangjin Yun · Seon-Ha Chae* · Sukmin Yoon[†]

경상대학교 건축도시토목공학부 및 공학연구원 · *K-water 융합연구원

Division of Architectural, Urban and Civil Engineering and Engineering Research Institute, Gyeongsang National University

*K-water Institute

(Received September 18, 2018; Revised November 6, 2018; Accepted November 26, 2018)

Abstract : Although multivariate outlier detection techniques have been actively studied in various fields, domestic studies on the water quality management of waterworks are relatively inadequate and have been performed in a small range. In this study, water quality dataset were collected from G_water treatment plants in South Korea and classified by the statistical correlation. For the groups with significant correlations, we compared and analyzed the outlier detection performance by applying distance and isolation forest techniques. For the group with insignificant correlation, we analyzed the outlier detection performance according to the change of machine learning instance after applying Isolation Forest method. As a result, the distance-based and Isolation Forest methods were able to effectively search global and local outliers in the water quality dataset. Furthermore Isolation Forest method is analyzed to search outliers in a wider range than the distance-based method. In the Isolation Forest method, the change of the outlier search performance according to the change of the machine learning amount is small.

Key Words : Water Quality Dataset, Outlier, Distance-Based Outlier Detection, Isolation Forest

요약 : 다변량 기반의 이상치 탐지 기법은 다양한 분야에서 활발히 연구되어 왔으나 상수도 수질관리 분야에 대한 국내 연구는 상대적으로 미흡하며, 작은 범위에서 수행되어왔다. 이에 본 연구에서는 국내 G_정수장을 대상으로 다항목의 수질자료를 수집하고 통계적 상관관계 유무에 따라 집단을 분류하였다. 그리고 상관관계가 유의미한 집단에 대해서는 거리기반 및 Isolation Forest 기법을 적용하여 기법별 이상치 탐색 성능을 상호 비교·분석하였으며, 상관관계가 유의미하지 못한 집단에 대해서는 Isolation Forest 기법만을 적용한 후 기계학습의 변화에 따른 이상치 탐색 성능을 분석하였다. 그 결과 Isolation Forest 기법이 거리기반 기법에 비해 보다 넓은 범위에서 이상치들을 탐색하는 것으로 분석되었으며, 기계학습량의 변화에 따른 이상치 탐색 성능의 변화는 미소한 것으로 나타났다.

주제어 : 다항목 수질자료, 이상치, 거리기반 이상치 탐색, Isolation Forest

1. 서론

상수도 시스템 내에서 수질관리를 목적으로 자동계측장비 등을 운영 하는 경우 가장 주요한 문제는 관측 및 수집되는 수질데이터들의 품질관리이다. 일반적으로 실시간으로 관측되는 수질항목들은 관측목적 및 대상지점에 따라 그 항목이 다양하며, 국내 대부분의 상수도 시스템에서의 수질측정은 분단위로 이루어짐에 따라 저장되는 데이터의 양 또한 매우 방대하다. 관측된 수질 원시자료(Raw data)는 수질측정장비 운영과정에서 발생하는 고장과 보정, 통신과정의 오류 및 시스템 내의 수리조건(Hydraulic condition)의 변화 등으로 인해 다양한 이상치(Outlier)를 포함할 가능성을 내포하고 있다. 따라서 수질관측과정에서 발생한 이상치들에 대한 객관적인 자료정리과정이 없이 원시자료를 즉각적으로 분석업무에 적용하는 경우 합리적인 결과를 도출하기란 사실상 불가능한 일이다.¹⁾

여기서 이상치란 주위 관측값들과는 너무 동떨어져 있음

으로 인해 전혀 다른 메커니즘으로 생성되었는가를 의심하게 하는 관측값으로 정의할 수 있으며, 이상치는 측정 중 잘못 기록되는 등의 실수(Mistake), 다른 자료와 섞여 발생되는 자료의 오염(Contamination), 그리고 측정항목이 가지고 있는 자료 고유의 변동성(Inherent variability)에 의해 발생되며 객관적인 자료 분석에 부정적인 영향을 미치므로 제거하는 것이 일반적이다.²⁾ 따라서 이상치 탐지 기법은 다양한 공학적 분야에서 적용이 요구되기 때문에 활발하게 연구되어져 왔으며, 단변량 분석(Univariate analysis)기반의 이상치 탐색과 다변량 분석(Multivariate analysis)기반의 이상치 탐색으로 대분할 수 있다.

단변량 분석 기반의 이상치 탐색기법 중 Discordancy Testing은 수질자료내의 이상치 탐지 목적으로 전통적으로 적용되어온 통계학적 접근법으로서, 이상치 검정을 위해 유도된 p-value가 한계값(Threshold or cut-off value)을 초과하는 경우 이상치로 분류하게 된다. Discordancy Testing 대표적인 접근법으로 Dixon's test, Discordance test, Rosner's

[†] Corresponding author E-mail: gnuysm@gmail.com Tel: 055-755-8707 Fax: 055-772-1799

test 및 Walsh's test 등이 있으나 다변량 기반의 이상치 탐지 기법은 극단치(Extreme value)와 이상치를 구분할 수 없고, 한계값 설정에 분석자의 주관이 개입됨으로 인해 신뢰도가 저하되는 단점이 있다. 또한 이상치 탐색과정에서 분석 변수간의 상관성(Correlation)을 고려할 수 없는 근본적인 한계를 내포하고 있다.^{3,4)}

반면 다변량 분석 기반의 이상치 탐색기법들은 분석 변수들의 상관성을 고려할 수 있으며, 이상치와 극단치를 구분할 수 있는 장점 등으로 인해 통계 및 비통계 분야에서 활발히 연구되고 있다. 다변량 분석 기반의 이상치 탐색기법은 분산 기반 접근, 거리기반 접근, 밀도기반 접근 및 분류모델기반 등으로 분류할 수 있다. 분산기반 이상치 탐색에서는 이상치들은 자료들의 집합에서 어떤 개체를 제거하게 되면 분산이 최소화 되는 개체들로 정의되며 평활화 요소(Smoothing factor)를 계산하여 이상치를 탐색한다.⁵⁾ 거리기반 접근의 이상치 탐색은 자료 내 개체들 사이의 유클리드 거리 또는 마할라노비스 거리(Mahalanobis distance)를 바탕으로 정상적인 자료들은 짧은 거리내에서 군집을 이루며 이상치들은 이러한 군집들로부터 멀리 떨어져 있다고 판단하게 된다.^{6,7)} 밀도기반 이상치 탐색에서는 자료개체들의 밀도를 나타내는 Local Outlier Factor (LOF)를 이용해 한 데이터 개체를 대상으로 그 점과 이웃한 개체들의 LOF를 비교하여 이상치를 판단하게 된다.⁸⁾ 그리고 분류모델기반 접근법은 자기연상 학습 신경망(Auto-associative neural network)과 같은 지도 학습을 통하여 학습 데이터를 통해 분류모델을 생성하고 이 분류 모델을 이용해 이상치와 정상치들을 탐색하는 기법이다.⁹⁾

이처럼 다변량 기반의 이상치 탐색기법은 다양한 분야에서 활발히 연구되어 왔으나 상수도 수질관리 분야에 대한 국내 연구는 상대적으로 미흡하며, 작은 범위에서 수행되어왔다. Heo 등¹⁰⁾은 실시간 측정된 수질자료에 대해 다변량 분석기법 중 주성분 분석을 수행하고, 호텔링의 T2-통계량을 활용하여 수질 이상 유무를 판단하는 시스템 개발한 바 있으며, Park 등¹¹⁾은 마할라노비스 거리를 이용한 수질자료의 이상치 탐색 시 극단값들의 영향을 최소화하기 위해 최소공분산행렬식으로부터 유도된 Robust 거리를 적용한 방법론을 제안한 바 있다. 하지만 호텔링의 T2-통계량, 마할라노비스 및 Robust 거리를 활용한 이상치 탐색은 모수적 통계(Parametric statistics) 기법에 기반함으로 인해 분석 자료의 통계적 분포 및 인자간의 상관성이 명확히 알려져 있지 않는 경우 적용하기 어려운 단점이 있다.

이에 본 연구에서는 모수기반의 다변량 이상치 탐지 기법의 단점을 보완하고자 분류모델 기반의 다변량 이상치 탐색기법인 Isolation Forest (iForest) 기법을 수질자료에 적용하였다. 여기서 iForest 기법은 기계학습 중 Decision tree를 활용한 분류모형 기반의 비모수적 이상치 탐색 기법으로서 대상 자료의 확률분포 및 인자간의 상관성이 불명확한 상황에서도 효과적으로 적용 가능할 것으로 예상된다. 그리고 본 연구에서 적용된 iForest 기법의 이상치 탐색 성능을 평가하기 위해 국내 G_정수장으로부터 pH, 수온(Temperature),

잔류염소(Residual chlorine), 탁도(Turbidity), 전기전도도(Electric conductivity) 5가지 항목으로 구성된 수질데이터를 수집하였으며 수질 자료의 해석을 위한 전처리 방안으로서 Isolation Forest 기법의 활용성 및 적용 가능성을 평가하는 것을 본 연구의 주요목적으로 하였다.

2. 이론적 배경

2.1. 거리기반의 이상치 탐색기법

마할라노비스 거리기반의 이상치 탐색기법은 다변량 가우시안 분포에 기초하며 자료들의 공분산 행렬과 이로부터 유도되는 마할라노비스 거리를 이용해 이상치들을 탐색한다. p 차원을 갖는 자료들의 벡터 $X=(x_1, ..., x_p)$ 에 대한 다변량 가우시안 분포의 확률밀도 함수(Probability density function)는 식 (1)과 같다.

$$f_X(x_1, ..., x_p) = \frac{1}{(\sqrt{2\pi})^p |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)\right) \quad (1)$$

여기서, μ 는 자료들의 평균벡터, Σ 는 공분산 행렬이다.

p 차원의 자료벡터 X 에 대한 마할라노비스 거리는 식 (2)와 같이 정의할 수 있으며, 마할라노비스 거리는 자유도가 p 인 χ^2 -분포를 따른 것으로 알려져 있다.

$$MD_i(X, \mu) = ((X_i - \mu)^T \Sigma^{-1}(X_i - \mu))^{1/2} \quad \text{for } i = 1, ..., n \quad (2)$$

그리고 마할라노비스 거리로부터 이상치를 탐색하기 위한 절사점(Cutoff point)은 자유도가 p 인 χ^2 -분포를 활용해 식 (3)과 같이 정의된다.

$$c = \sqrt{\chi_{p; 1-\alpha}^2} \quad (3)$$

여기서, α 는 유의수준(Significance level)이다.

Acuna 등¹²⁾은 다변량 자료에서 다량의 이상치가 분포하는 경우 가면화 현상(Masking phenomenon) 및 수렁 현상(Swamping phenomenon)이 발생되며 이로 인해 이상치를 탐색하기 위한 성능이 저하된다고 밝힌 바 있다. 여기서, 가면화 현상이란 일부 이상치가 다른 이상치에 의해 정상치로 판정되는 현상이며, 수렁 현상이란 이상치의 영향에 의해 정상치가 이상치로 판단되는 현상을 말한다.

마할라노비스 거리 또한 자료 내 분포하는 잠재적 이상치들과 극값들의 영향으로 인해 추정해야 할 모수들이 강건하지 못한 경우 효율적이지 못하다는 문제점을 내포하고

있다. 따라서 최근에는 식 (4)에 나타난 것과 같이 이상값과 극단값들의 영향을 최소화하기 위한 Robust 거리가 적용되고 있다.

$$RD(X, T(X)) = \sqrt{(X - T(X))C(X)^{-1}(X - T(X))^T} \quad (4)$$

여기서 $T(X)$ 는 자료벡터 X 의 Robust 위치, $C(X)$ 는 Robust 산포행렬이다.

일반적으로 Robust 추정치를 사용하는 방법으로는 MCD (Minimum covariance determinant)에 의한 방법, MVE (Minimum volume ellipsoid)에 의한 방법, OGK (Orthogonalized gnanadesikan-kettenring)에 의한 방법 등이 있으며,¹³⁻¹⁷⁾ 본 연구에서는 Robust 거리를 산정하기 위해 MCD에 의한 방법을 적용하였다. MCD 추정량은 N 개의 자료에서 공분산 행렬이 가장 최소가 되는 부분집합을 통해 계산된다. 부분집합은 최소 $[N/2]+1$ 개의 자료를 포함해야 하며, 계산된 평균벡터를 MCD 위치추정량(Location estimator), 공분산 행렬을 MCD 산포행렬(Scatter estimator)라 정의한다. 따라서 MCD 추정량을 적용한 Robust 거리는 다음 식 (5)와 같다.

$$RD_i(X, \mu_{MCD}) = \sqrt{(X_i - \mu_{MCD})\Sigma_{MCD}^{-1}(X_i - \mu_{MCD})^T} \quad (5)$$

Robust 거리는 마할라노비스 거리와 동일하게 자유도가 p 인 χ^2 -분포 따르며 식 (3)에서 나타난 절사점을 활용하여 이상치를 판단할 수 있다.

2.2. iForest 기법

iForest 기법은 자료들의 거리 또는 밀도에 의존하지 않고 기계학습 방법론인 Decision tree를 이용해 분류모형을 생성하고 개체고립을 통해 이상치를 탐지하게 된다.^{18,19)}

Fig. 1은 2차원 공간에서 iForest 기법의 이상치 탐색을 위한 개체 고립사례를 나타낸 것이다.⁹⁾ iForest 기법은 임의의 기준으로 무작위로 공간을 분할하며, 정상치로 분류되는 개체 x_i 를 고립시키기 위해서는 일반적으로 더 많은 공간분할을 수행해야 한다. 하지만 이상치로 판단되는 개체인 x_o

Fig. 1. Isolating a normal point x_i and abnormal x_o .⁹⁾

의 경우 더 적은 공간분할만으로 고립시킬 수 있다. 이러한 공간분할은 Fig. 2에서 나타낸 의사결정나무 형태인 Isolation tree로 표현할 수 있다.¹⁹⁾

이상치는 특성상 고립에 취약하기 때문에 Isolation tree의 뿌리에 가깝게 고립될 가능성이 높아 상단부에 위치하며, 정상치는 Isolation tree의 하단부까지 깊게 위치하며, 이러한 경로 길이를 기준으로 이상치와 정상치를 분류할 수 있게 된다. 이상치 탐색을 위한 공간분할은 무작위로 생성되며 예상 경로 길이를 계산하기 위해 Isolation tree를 반복 생성하여 평균 경로 길이를 계산하며, 이상치인 개체의 평균경로 길이는 정상개체의 평균경로 길이에 비해 짧게 나타나게 된다. 그리고 n 회의 반복학습을 통해 계산된 개체들의 평균경로 길이들로부터 이상치를 여부를 판단하기 위해서는 식 (6)에서 나타낸 정규화된 이상점수가 필요하다.

$$\begin{cases} c(n) = 2H(n-1) - (2(n-1))/n \\ s(x, n) = 2 \frac{E(h(x))}{c(n)} \end{cases} \quad (6)$$

여기서, $s(x, n)$ 은 대상개체의 이상점수, $h(x)$ 경로길이, $H(i)$ 는 Harmonic number로 $\ln(i) + 0.57721\dots$ (Euler's constant)에 의해 추정된다.

식 (6)로부터 정규화된 이상점수의 범위는 0~1로서 일반적으로 이상점수가 1에 매우 가깝게 반환되면 개체는 확실한 이상치로 판단이 가능하며, 0.5보다 작을 경우 정상치로 판단이 가능하다.

Fig. 2. Description of isolation tree.¹⁹⁾

3. 연구방법

본 연구에서는 다항목 수질자료에 대한 iForest 기법의 이상치 탐색 성능을 분석하고자 거리기반 이상치 탐색기법과의 비교를 수행하였다. 거리기반 이상치 탐색에서는 마할라노비스 거리 및 Robust 거리가 적용되었으며, 각 기법별 이상치 탐색 성능을 상호 비교·분석하기 위해 2013년 1월 1일~2015년 7월 31일 기간 관측된 시단위 5개 항목(pH, 잔류염소, 수온, 탁도 및 전기전도도)의 수질자료를 국내 G_정수장으로부터 수집하였다. 그리고 Fig. 3은 5개 수질항목들의 시계열분포를 나타낸 것이며, 수질항목별 통계적 특성은 Table 1에 요약하였다.

앞 절에서 설명한 것과 같이 마할라노비스 거리 및 Robust 거리 기반의 이상치 탐색 기법은 분석인자들 사이의 통계적

Fig. 3. Time series plot of water quality data from G_WTP.

Table 1. Statistical characteristics of water quality data from G_WTP

Water quality	Mean	Std.
pH	7.19	0.36
Residual chlorine (mg/L)	0.59	0.08
Temperature (°C)	13.68	7.98
Turbidity (NTU)	0.05	0.01
Electric conductivity (μS/cm)	191.19	39.86

상관관계가 유의한 경우에만 적용 가능하다. 따라서 본 연구에서는 5개 수질항목들에 대해 상관분석을 수행하여 수질항목들을 분류하였다. 그리고 분류된 수질항목중 상관성이 유의한 그룹에 대해서는 iForest 및 거리기반 이상치 탐색 기법을 적용하여 이상치 탐색성능을 상호 비교하였으며, 상관성을 만족하지 못한 그룹에 대해서는 iForest 기법만을 적용한 후 기계학습 변화에 따른 이상치 탐색성능을 분석하였다.

아울러 각 기법별로 식별된 이상치에 대한 최종적인 판단을 위해서는 해당 정수장에 수집된 수질자료 이외에도 정수 공정, 계측기 관리 등에 관한 상세 운영기록이 요구되나 해당 자료들을 수집할 수 없는 현실적 한계로 인해 본 연구에서는 식별된 이상치들을 ‘이상치 후보군(Outlier candidates)’으로서 정의하였다.

4. 수질자료의 이상치 탐색 결과

연구대상지점으로부터 수집된 5개 수질항목들을 상관성 유무에 따라 분류하기 위해 피어슨 상관계수를 산정하였고, 그 결과는 Table 2에 요약하였다.

Table 2에 요약한 것과 같이 pH-수온의 상관계수는 -0.38로서 약한 음의 상관관계를 나타낸 반면 잔류염소-탁도-전기전도의 상관계수들은 -0.02~0.20로서 매우 약한 상관관계를 나타내었다. 따라서 본 연구에서는 유의한 통계적 상관관계를 나타내는 pH-수온을 A집단 그리고 상관관계가 매우 약한 잔류염소-탁도-전기전도를 B집단으로 분류하였다.

유의미한 통계적 상관관계가 나타난 A집단에 대한 예비조사를 위해 Z-score를 산정하였으며, 항목별 Z-score의 분포는 Fig. 4에 나타내었다.

Table 2. Correlation analysis result of water quality data from G_WTP

	pH	Residual chlorine	Temperature	Turbidity	Electric conductivity
pH	1.00	0.20	-0.38	0.11	-0.02
Residual chlorine	0.20	1.00	-0.01	0.02	-0.05
Temperature			1.00	0.13	0.15
Turbidity				1.00	-0.02
Electric conductivity					1.00

Fig. 4. Scatter plot of Z-score((a) pH, (b) Temperature).

Fig. 5. Scatter plot of standardized pH and temperature.

Z-score는 단변량 이상치를 식별하기 위한 기본적인 접근법으로 상·하한 이상치의 경계값으로 ± 3 ($p\text{-value} < 0.0014$)을 설정하는 경우 pH는 79개(0.35%), 수온은 0개의 이상치 후보군이 탐색되었다.

하지만 Fig. 5에서 나타난 집단A의 Z-score 산점도를 살펴보면 전역 이상치(Global outlier)로 의심되는 다량의 관측값들과 더불어 지역 이상치(Local outlier)로 의심되는 관측값들의 군집들이 정상 관측치 군집과 인접한 위치에서 분포하고 있는 것을 알 수 있다. 이처럼 다변량 자료에서 나타나는 통계적 상관관계를 고려하지 않고 개개의 자료들을 대상으로 단변량 이상치 탐색만을 수행하는 경우 실제 분포하는 지역 및 전역 이상치들을 효과적으로 탐지하지 못하는 오류가 발생할 수 있게 된다.

본 연구에서는 유의미한 통계적 상관관계가 나타나는 이변량 집단A의 이상치 탐색을 수행하기 위해 거리기반 및 iForest 기법을 적용하였다. 거리기반의 이상치 탐색은 마할라노비스 거리 및 Robust 거리를 이용해 수행되었으며, 이상치 판단을 위한 절사점은 χ^2 -분포의 유의수준 0.025에 해당하는 $C = \sqrt{\chi^2_{2,0.975}} = 2.176$ 로 설정하였다. 그리고 Table 3은 마할라노비스 거리 및 Robust 거리를 계산하기 위해 추정된 위치매개변수, 공분산 행렬 및 각 행렬식을 요약한 것이다. Robust 매개변수 추정은 MCD기법, 샘플의 크기는 전체 자료의 크기(N)의 75%를 적용하였다.

Table 3에 요약한 것과 같이 MCD기법을 통해 추정된 공

Table 3. Estimation results of parameters for mahalanobis & robust distance

Distance	Mean (pH, Temp.)	Covariance (pH-Temp.)		det (Cov.)
Mahalanobis	(7.19, 13.68)	0.13	-1.11	7.05
		-1.11	63.67	
Robust	(7.21, 13.75)	0.06	-2.12	1.41
		-2.12	98.43

Fig. 6. Outlier detection results of group A ((a) mahalanobis distance, (b) robust distance, (c) iForest).

분산 행렬식은 원시자료의 공분산 행렬식에 비해 약 80%정도 감소하는 것으로 나타났다. 그리고 iForest를 활용한 이상치 탐색을 위해서는 기계학습의 횟수 n , 이상점수 s 의 설정이 필요하다. 일반적으로 이상치 탐색을 위해 iForest기법을 적용하는 경우 기계학습의 횟수 $n = 50 \sim 100$, 이상점수 $s > 0.5$ 의 조건이 적합한 것으로 알려져 있으며, 본 연구에서는 기계학습의 횟수 $n = 75$, 이상점수 $s = 0.6$ 의 조건을 적용하였다.

Fig. 6은 각 기법들에 의해 탐색된 이상치 후보군들의 분포를 나타낸 것이다. Fig. 6(a)는 마할라노비스 거리를 활용한 이상치 탐색결과를 나타낸 것으로 Z-score 경계 내부에서 분포하는 전역 이상치 및 자료 상하부에서 군집으로 분포하는 지역 이상치들을 또한 이상치 후보군으로서 식별되는 것을 알 수 있다. 그리고 Fig. 6(b)는 Robust 거리를 활용한 이상치 탐색결과를 나타낸 것으로 Z-score 경계 내부에서 분포하는 전역 이상치 및 지역 이상치들이 보다 뚜렷하게 식별되는 것을 알 수 있다. Fig. 6(c)는 iForest 활용한 이상치 탐색결과를 나타낸 것으로 Z-score 경계 내부에서 분포하는 전역·지역 이상치 및 자료 분포의 양쪽 꼬리부분에서 나타나는 일부 극값들을 이상치 후보군으로 식별하는 것을 알 수 있다.

본 연구에서는 각 기법별 이상치 탐색 성능과 이에 따른

Table 4. Results of linear regression analysis

Group	# of outlier candidates	Correlation coefficient	Regression model ; $y = \beta + \alpha x$		
			β	α	R^2
Raw data	-	-0.38	7.43	-0.017	0.15
MD	734 (3.2%)	-0.81	7.51	-0.021	0.67
RD	1,332 (5.9%)	-0.85	7.52	-0.022	0.72
iForest	4,570 (20.2%)	-0.86	7.51	-0.021	0.74

통계적 효과를 정량적으로 검증하기 위해 원시자료 및 이상치들이 제거된 자료들에 대해 선형회귀분석을 수행하고 설명력의 정도를 나타내는 결정계수(R^2)의 변화를 분석하였다.

Table 4는 원시자료 및 이상치가 제거된 자료들에 대한 선형회귀분석 결과를 요약한 것으로 집단A의 원시 상관관계수는 -0.38, 각 기법들을 통해 이상치 후보군들을 제거한 집단A의 상관관계수는 -0.81~-0.86으로서 통계적 상관성이 보다 뚜렷하게 나타나는 것을 알 수 있다. 그리고 원시자료 집단을 대상으로 수행된 선형회귀분석의 결정계수는 0.15로서 집단A의 선형관계에 대한 설명력이 유의미하지 못한 반면, 이상치 후보군을 제거한 집단A의 결정계수들은 0.67~0.74로서 선형관계에 대한 통계적 설명력이 유의미한 것으로 나타났다.

로 나타났다.

잔류염소-탁도-전기전도로 구성된 집단B의 피어슨 상관계수는 -0.02~0.20로서 매우 약한 상관관계를 나타내었다. 따라서 집단B에 대해서는 iForest 기법만을 적용하여 이상치 탐색을 수행하였으며, 기계학습 반복횟수(n) 따른 이상치 탐색 성능 변화를 분석하였다. 기계학습의 반복횟수의 일반적 적용범위는 $n = 50 \sim 100$ 이며 본 연구에서는 $n = 50$, $n = 75$, $n = 100$ 의 3가지 조건을 이용해 이상치 탐색을 수행하였다.

Fig. 7(a)는 집단B의 원시자료 분포를 나타낸 것으로 정상군집을 중심으로 전역 이상치 및 지역 이상치 군집으로 의심되는 관측값들이 자료 전반에 분포하고 있는 것을 알 수 있다. 그리고 Fig. 7(b.1)~7(d.2)는 iForest ($s = 0.6$) 기법의 기계학습 반복횟수 변화에 따른 이상치 탐색결과를 나타낸 것으로서, 원시자료 내에 분포하고 있는 전역 이상치 및 지역 이상치 군집들이 모두 효과적으로 탐색되는 것을 알 수 있다. 그리고 기계학습의 반복횟수 따른 이상치 후보군의 탐색수는 $n = 50$ 에서 2,177개(9.63%), $n = 75$ 에서 2,057개(9.10%) 그리고 $n = 100$ 에서 2,161개(9.56%)로서 기계학습의 반복횟수의 변화에 따른 iForest 기법의 이상치 탐색 성능 변화는 미소한 것으로 나타났다.

Fig. 7. Outlier detection results of group B.

5. 결론

다변량 기반의 이상치 탐지기법은 다양한 분야에서 활발히 연구되어 왔으나 상수도 수질관리 분야에 대한 국내 연구는 상대적으로 미흡하며, 작은 범위에서 수행되어왔다. 이에 본 연구에서는 국내 G_정수장을 대상으로 다항목의 수질자료를 수집하고 통계적 상관관계 유무에 따라 집단을 분류하였다. 그리고 상관관계가 유의미한 집단에 대해서는 거리기반 및 iForest 기법을 적용하여 기법별 이상치 탐색 성능을 상호 비교·분석하였으며, 상관관계가 유의미하지 못한 집단에 대해서는 iForest 기법만을 적용한 후 기계학습의 변화에 따른 이상치 탐색 성능을 분석하였다. 그 결과들을 요약하면 다음과 같다.

1) 유의미한 통계적 상관성을 나타낸 집단A(pH-수온)에 대해 Z-score를 적용해 이상치들을 탐색한 결과 다변량 선형관계에서 분포하는 전역 및 지역이상치 군집들을 효과적으로 탐색할 수 없었다. 반면, 거리기반 및 iForest 기법을 적용한 다변량 이상치 탐색에서는 단변량 이상치 탐색의 경계값 내부에서 발생하는 전역 이상치 및 지역이상치 군집들을 효과적으로 탐색할 수 있었다.

2) 거리기반의 이상치 탐색을 위해 마할라노비스 거리 및 Robust 거리를 적용하였다. MCD 추정에 의해 공분산 행렬식은 80% 감소하였으며, 이상치 탐색수는 약 1.8배 이상 증가하는 것으로 나타났다. 그리고 iForest 기법을 적용하는 경우 Robust 거리기법에 비해 약 3.4배 이상 높은 이상치를 탐색할 수 있었으며, 특히 원시자료 분포의 양쪽 꼬리에 위치한 일부의 극값들 또한 이상치로 탐색하는 것으로 나타났다.

3) 통계적 상관성이 낮은 집단B(잔류염소-탁도-전기전도도)에 대해 iForest 기법만을 적용한 후 기계학습의 반복횟수 따른 이상치 탐색 성능을 분석하였다. 기계학습의 반복횟수의 일반적 적용범위는 $n = 50 \sim 100$ 이며 본 연구에서는 $n = 50$, $n = 75$, $n = 100$ 의 3가지 조건을 이용해 이상치 탐색을 수행한 결과 $n = 50$ 에서 9.63%, $n = 75$ 에서 9.10% 그리고 $n = 100$ 에서 9.56%의 이상치가 탐색되었으며, iForest 기법에서 기계학습의 반복횟수 변화에 따른 이상치 탐색 성능의 변화는 미소한 것으로 나타났다.

상수도시스템 내에서 관측되는 수질자료들로부터 통계적 분석을 통해 합리적인 의사결정을 도출하기 위해서는 원시자료에 대한 이상치 탐색 작업이 필수적이라고 할 수 있다. 본 연구의 주요결과들에서 제시된 것과 같이 기존의 단변량 기반의 이상치 탐색기법은 상관관계가 존재하는 수질항목들에 대해서는 적용성의 한계가 명확한 만큼 다변량 기반의 이상치 탐색이 보다 더 유효할 것으로 판단된다. 아울러 관측되는 수질자료들의 통계적 분포 및 상관관계가 명확하지 않는 경우가 일반적이라는 상황을 고려할 때 본 연

구에서 제안한 기계학습 기반의 이상치 탐색기법인 iForest 기법은 효과적인 이상치 탐색 도구로서 적용 가능할 것으로 기대된다.

Acknowledgement

본 연구는 환경부 물관리연구사업의 연구비 지원(과제번호: 16AWMP-B113766-01)에 의해 수행되었습니다.

KSEE

References

1. Robinson, R. B., Cox, C. D. and Odom, K., "Identifying outliers in correlated water quality data," *J. Environ. Eng.*, **131**(4), 651~657(2005).
2. Hawkins, D. M., Identification of outlier, Chapman and Hall, London, (1980).
3. Barnett, V. and Lewis, T., Outliers in statistical data, 3rd ed, John Wiley & Sons. (1994).
4. U. S. EPA, Data Quality Assessment : Statistical Methods for Practitioners, EPA QA/G-9S, (2006).
5. Arning, A., Agrawal, R. and Raghavan, P., "A linear method for deviation detection in large multidimensional databases," in Proceedings Intm Confm on Knowledge Discovery and Data Mining, Portland, OR, (1996).
6. Filzmoser, P., Reimann, C. and Garrett, R. G., "Multivariate outlier detection in exploration geochemistry," Technical report TS 03-5, Department of Statistics, Vienna University of Technology, Austria(2003).
7. Filzmoser, P., "A multivariate outlier detection method," In Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling, Minsk, Belarus, **1**, 18~22(2004).
8. Breunig, M. M., Kriegel, H. P., Ng, R. T. and Sander, J., "LOF: identifying density-based local outliers," in Proceedings of 2000 ACM SIGMOD International Conference on Management of Data, ACM Press, pp. 93~104(2000).
9. Liu, F. T., Ting, K. M. and Zhou, Z. H., "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **6**(1), 31~39(2012).
10. Heo, T. Y., Jeon, H. B., Park, S. M. and Lee, Y. J., "Development of Real-Time Water Quality Abnormality Warning System for Using Multivariate Statistical Method," *J. Korean Soc. Civil Eng.*, **37**(3), 137~144(2015).
11. Park, N. S., Kim, S. S., Chae, S. and Yoon, S., "Introducing New Outlier Detection Method Using Robust Statistical Distance in Water Quality Data," in Proceedings of 7th IWA-ASPIRE Conference, (2017).
12. Acuna, E. and Rodriguez, C., "The treatment of missing values and its effect on classifier accuracy," Classification, Clustering, and Data Mining Applications, Banks, D., McMorris, F. R., Arabie, P. and Gaul, W.(Eds), Springer, Berlin,

- Heidelberg, pp. 639~647(2004).
13. Rousseeuw, P. J., "Least median of squares regression," *J. Am. Statistical Assoc.*, **79**(388), 871~880(1984).
 14. Rousseeuw, P. J. and Yohi, V. J., "Robust regression by means of S-estimators," *Robust and Nonlinear Time Series Analysis*, Franke, J., Hardle, W. and Martin, D. (Eds.), Springer, New York, pp. 256~272(1984).
 15. Rousseeuw, P. J. and Leroy, A. M., *Robust regression and outlier detection*, John Wiley & Sons., NewYork, (1987).
 16. Rousseeuw, P. J. and Driessen, K. V., "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, **41**(3), 212~223(1999).
 17. Maronna, R. A. and Zamar, R. H., "Robust multivariate estimates for high dimensional datasets," *Technometrics*, **44**(4), 307~317(2002).
 18. Liu, F. T., Ting, K. M., and Zhou, Z. H., "Isolation Forest," in *Proceedings of 2008 Eighth IEEE International Conference on Data Mining (ICDM)*, pp. 413~422(2008).
 19. Chen, W. R., Yun, Y. H., Wen, M., Lu, H. M., Zhang, Z. M., and Liang, Y. Z., "Representative subset selection and outlier detection via isolation forest," *Anal. Methods*, **8**(39), 7225~7231(2016).