기계학습에 기반한 댐 수위 이상 데이터 탐지

방수일*, 이도길**
*고려대학교 컴퓨터정보통신대학원 빅데이터융합학과
**고려대학교 민족문화연구원
banggyber@korear.ac.kr, motdg@korear.ac.kr

Detection of Abnormal Dam Water Level Data Based on Machine Learning

Suil Bang*, Do-Gil Lee**
*Dept. of Big Data Convergence, Korea University
**Research Institute of Korean Studies, Korea University

요 약

K-water에서는 다목적댐의 관리를 위해 실시간으로 댐수위, 하천 수위 및 강우량 등을 계측하고 있으며, 계측된 값들은 댐을 효과적으로 운영하는데 필요한 데이터로 활용되고 있다. 특히 댐수위 이상 데이터를 탐지하지 못한 채 그대로 사용할 경우 댐의 방류 시기와 방류량 등을 결정하는 중요한 의사결정을 그르칠 수 있으므로 이를 신속히 탐지하는 것이 매우 중요하다. 현재의 자동화된 이상 데이터 탐지방법 중 하나는 현재 데이터가 최댓값과 최솟값을 초과할 때, 다른 하나는 현재 데이터와 일정 시간 동안의 평균값 간의 차이가 관리자가 정한 특정 값을 벗어났을 때를 기준으로 삼고 있다. 전자는 상한과 하한의 초과 여부만 판단하므로 탐지가 쉬우나 정상범위 내에서 발생한 이상 데이터는 탐지가 불가하다. 후자는 관리자의 경험을 통해 판단 조건을 정하기 때문에 객관성이 결여되는 문제가 있다. 특히 방류와 강우가 복합적으로 댐수위에 영향을 미치는 홍수기에 관리자의 경험에 기초한 이상 데이터 판별은 신뢰성의 문제가 있을 수 있다. 따라서 본 연구에서는 기계학습을 최초로 적용하여 이상 데이터를 탐지하고자 하였다. 댐수위, 누적강우량 및 누적방류량 데이터와 댐수위데이터를 가공하여 생성한 댐수위차, 댐수위차평균, 댐수위평균 등 자질들의 다양한 조합을 만든 후 이를 Random Forest, SVM, AdaptiveBoost 및 다층퍼셉트론(MLP) 등과 같은 여러 가지 기계학습모델 등을 통해 이상 데이터를 판별하는 실험(분류)을 하였다. 실험결과 댐수위, 댐수위차, 댐수위-댐수위평균, 누적강우량, 누적방류량 및 댐수위차평균을 사용하였을 때 MLP에서 가장 우수한 성능을 보였다. 이 연구를 통해서 댐수위 이상 데이터를 기계학습의 분류기능을 통해 효과적으로 탐지할 수 있다는 것과 모델의 성능은 실험에 사용한 자질의 수뿐 아니라 자질의 종류에도 큰 영향을 받는다는 것을 알 수 있었다.

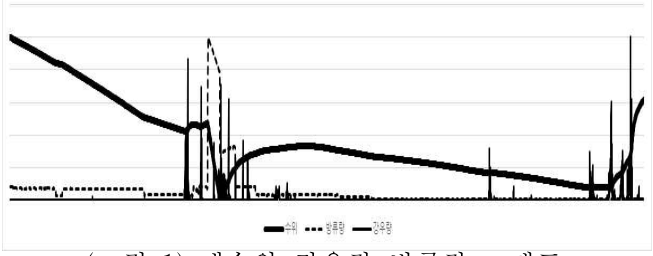
Keyword : 이상치 검출, Machine Learning, Random Forest, SVM, AdaptiveBoost, Multi-layer Perceptron

1. 서론

수문 데이터는 수자원 조사 및 댐 운영관리에 필요한 기초정보로써 주로 현대화된 자동관측 시스템으로부터 취득된다. 댐 운영 측면에서는 댐 상류에서 측정되는 강수량과 하천 수위, 댐 자체에서 측정하는 댐수위 및 방류량 등이 의사결정을 위한 주요 데이터이다

하천의 홍수피해를 미리 방지하기 위해서는 강우유출 현상의 해석이나, 과거 수문 자료를 이용한 통계해석에 의한 합리적인 하천 구조물의 설계 및 홍수 예측기술이 필요하며, 이의 근거로 수문 데이터가 필요하다. [1]

수문 데이터는 댐 주변의 관측소들로부터 많은 양의 데이터가 실시간으로 수집되는 스트리밍 형태로 이 데이터들은 센서 데이터의 전형적인 특징을 가지고 있고 통신시스템을 통해 전송된 후 일정 간격으로 데이터베이스에 저장 및 관리된다. 현대화된 자동관측 시스템을 통해 전송되더라도 관측기기 및 통신장비의 결함이나 데이터 처리 과정에서의 문제로 인해 이상 데이터가 발생할 수 있는데 평상시에는 큰 문제가 되지 않지만, 홍수가 발생하는 시기에 발생한 댐 수위 이상 데이터는 댐의 방류 시기나 방류량을 결정하는 의사결정에 심각한 영향을 초래할 수 있다.



(그림 1) 댐수위-강우량-방류량 그래프

(그림 1)은 댐수위-강우량-방류량 그래프이다. 굵은 실선이 댐 수위, 가는 실선이 강우량, 점선이 방류량을 나타낸다. (그림 1)의 오른쪽 부분을 보면 강우가 발생한 뒤 수위가 증가하는 것을 알 수 있고 왼쪽 부분을 보면 방류량 증가에 따라 댐 수위가 급격히 감소했다가 강우의 영향으로 다시 증가하는 것을 알 수 있다.

다목적댐의 효과적인 운영을 위해 이상 데이터 탐지는 매우 중요하다. 그러나 수문 데이터는 실시간 전송되는 데이터이기 때문에 관리자가 지속해서 이상 유무를 판별하는 것은 불가능하다.

따라서 현업에서 자동화된 방법으로 댐 수위 이상 데이터를 탐지하는 방법은 두 가지를 사용한다. 첫 번째 방법은 댐에서 측정될 수 있는 수위 데이터의 최댓값과 최솟값을 이용하는 방법이다. 일반적으로 이 두 값은 자주 발생하지 않을 뿐 아니라 발생하더라도 쉽게 탐지할 수 있다. 두 번째 방법은 현재 데이터와 일정 시간 동안의 평균 수위 데이터의 차이가 관리자가 정해 놓은 일정 값만큼 벗어날 때 이상 데이터로 검출하는 방법이다. 예를 들어 30분 동안의 댐 수위 평균과 현재 데이터가 10cm 이상 차이가 나면 이상 데이터로 판단하는 식이다.

댐 수위가 잘 변하지 않는 기간에는 이 방법은 크게 문제가 없다. 그러나 (그림 1)과 같이 방류와 강우가 동시에 댐수위에 영향을 미치는 시기에는 이러한 방법은 신뢰도가 떨어질 수밖에 없다.

따라서 본 연구에서는 그간 현업에서 이상 데이터 탐지에 활용하지 않았던 기계학습을 통해 이상 데이터를 탐지하는 실험을 해보았다.

2. 분석방법

2-1. 학습데이터

원시 수문 데이터는 매 1분 단위로 저장되는데 이 데이터는 분석에 사용하지 않으므로 10분 데이터[1)]를 활용하였다.

분석에 사용된 데이터는 횡성댐의 10분 데이터를 활용하였다. 2017년 7월~8월 2달 동안의 10분 데이터 8,640개 중 강우량과 방류량의 영향으로 수위 변화가 잦았던 구간을 선정하여 일차적으로 약 1,800여 개를 추출하였으며, 2차 추출을 통해 약 20%인 364개를 다시 추출하였다. 추출 시 최대한 고르게 추출될 수 있도록 데이터를 추출하였다.

데이터베이스에서 추출한 데이터는 모두 정상 데이터이므로 학습과 실험을 위해 일부 데이터를 이상 데이터로 변경하였다. 변경된 데이터는 약 10% 정도로 최대한 객관성을 확보할 수 있도록 여러 번 검증을 거쳤다.

실험에 적용한 데이터 예시는 <표 1>과 같다.

<표 1> 실험에 적용한 데이터 예시

댐 수위	댐수위 차	댐수위 평균	누적 방류량	누적강우량	판단
17080	1	17079.00	4.09	4.67	정상
17394	4	17390.00	48.18	14.26	이상
17399	2	17395.33	49.26	14.25	정상
17139	3	17134.67	21.82	10.88	이상
17683	2	17676.00	23.91	13.31	이상

2-2. 자질 구성

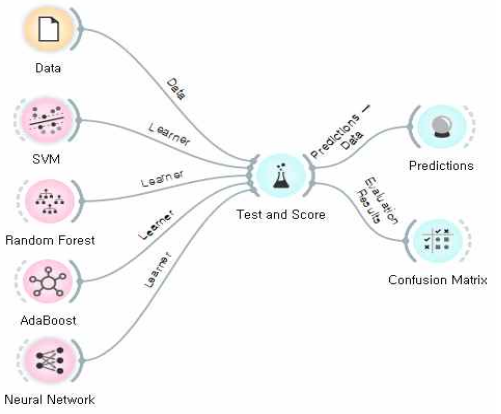
실험을 위해 <표 2>과 같은 자질을 사용하였다.

<표 2> 실험에 사용된 자질

번호	명칭	단위	설명
F1	댐 수위	cm	현재 댐 수위
F2	댐수위 차	cm	현재수위 - 이전수위
F3	댐수위 평균	cm	댐 수위의 평균 (30분 평균, 현재 제외)
F4	누적 방류량	m^3/s	30분간 누적 방류량 (현재 값 포함)
F5	누적 강우량	mm	3시간 지연된 6시간 누적강우량

2-3. 모델 구성 및 평가방법

실험에 사용한 기계학습 도구는 Orange Python v2.6을 사용하였다. Orange Python은 (그림 2)와 같이 그래픽 요소를 이용하여 쉽게 기계학습 모델을 구성하고 평가를 할 수 있는 프로그램이다.



(그림 2) Orange Python

1) 1분 데이터 10개를 평균한 데이터

Random Forest, SVM, Adaptive Boost 및 다층 퍼셉트론(Multi-layer Perceptron, MLP)을 사용하여 모델을 구성하고 실험을 시행하였다. SVM의 Kernel은 RBF를 사용하였고 MLP는 1개의 은닉층, 뉴런수는 20-30개, 활성화 함수를 Sigmoid, 최적화 알고리즘은 L-BFGS-B를 사용하였다.

실험에 사용한 데이터 세트는 정상 데이터가 훨씬 많은 불균형 데이터 세트이다. 또한, 본 실험은 이상 데이터를 얼마나 잘 예측하느냐가 중요하므로 성능 평가는 <표 3>과 같이 혼동행렬을 활용하여 정밀도(Precision)와 재현율(Recall), F1스코어를 모두 평가하되 재현율에 좀 더 비중을 두고 평가하였다. [2]

<표 3> 혼동행렬 및 모델의 성능평가

		예측	
		이상	정상
실제	이상	TP	FN
	정상	FP	TN
정밀도 : TP/(TP+FP). 재현율 : TP/(TP+FN) F1스코어 : 2*(정밀도*재현율)/(정밀도+재현율)			

전체 모델에서 K겹 교차검증(K=10)을 시행하였으며, 각 Fold는 층화추출법[2]을 사용하여 구성하였다.

교차검증 시행 후 모델별 성능을 비교하였다.

3. 평가 결과

3-1. 1차 실험 및 평가

다양한 실험을 위해 <표 4>와 같이 자질들의 조합을 구성하여 실험을 시행하였다.

<표 4> 자질 조합의 구성(1차)

구분	F1	F2	F3	F4	F5
조합1	O	X	X	O	O
조합2	O	X	O	X	X
조합3	O	X	O	O	O
조합4	O	O	X	X	X
조합5	O	O	X	O	O
조합6	O	O	O	X	X
조합7	O	O	O	O	O

2) 표본 추출 시 모집단을 먼저 중복되지 않도록 층으로 나눈 다음 각 층에서 표본을 추출하는 방법

<표 5> 기계학습 모델별 실험 결과(%)

조합	평가지표	SVM	Random Forest	Ada Boost	MLP
1	F1	0.00	16.39	21.21	28.00
	P	0.00	18.52	21.88	43.75
	R	0.00	14.72	20.59	20.59
2	F1	0.00	23.33	21.82	19.05
	P	0.00	26.92	17.65	50.00
	R	0.00	20.59	28.57	11.76
3	F1	0.00	26.23	20.59	34.48
	P	0.00	29.63	20.59	41.67
	R	0.00	23.53	20.59	29.41
4	F1	68.75	69.70	74.19	74.19
	P	73.33	71.88	82.14	82.14
	R	64.71	67.65	67.65	67.65
5	F1	70.97	59.79	70.59	73.50
	P	78.57	60.60	70.59	73.50
	R	64.71	58.82	70.59	73.50
6	F1	68.75	67.61	75.00	77.42
	P	73.33	64.86	80.00	85.71
	R	64.71	70.59	70.59	70.59
7	**F1**	**68.85**	**72.22**	**70.59**	**79.99**
	P	**77.77**	**68.42**	**70.59**	**77.77**
	R	**61.64**	**76.47**	**70.59**	**82.36**

* F1(F1스코어), P(정밀도), R(재현율)

<표 5>의 실험 결과를 통해 모델 성능에 유의미한 변화가 있는 부분을 찾아보면 조합 1-3, 조합 4-7의 두 개의 그룹으로 나눌 수 있다.

<표 4>를 보면 F2(댐수위차)가 조합에 포함된 경우와 그렇지 않은 경우로 구분할 수 있다. 실험 결과를 바탕으로 F2(댐수위차)가 모델의 성능에 큰 영향을 미치는 자질임을 알 수 있다. 1차 실험에서는 자질 조합 7을 사용하였을 때 MLP가 가장 높은 성능을 보였다.

3-2. 자질 추가 및 평가

이번 실험에서는 자질 추가에 따른 모델들의 성능변화를 실험하기 위해 '댐 수위 차 평균'을 추가하였다.

<표 6> 추가자질

번호	명칭	단위	설명
F6	댐수위차평균	cm	댐수위차 30분 평균 (현재 데이터 제외)

이 자질을 추가한 이유는 홍수기와 같이 수위가 크게 변하는 상황에서 수위 변화의 추이를 이상 데이터 탐지에 반영하기 위함이다.

<표 7>과 같이 자질들의 조합을 구성하고 실험한 결과는 <표 8>과 같다.

<표 7> 자질 조합의 구성(2차)

구분	F1	F2	F3	F4	F5	F6
조합8	O	X	X	O	O	O
조합9	O	X	O	O	O	O
조합10	O	O	X	O	O	O
조합11	O	O	O	O	O	O

<표 8> 자질 추가에 따른 평가 결과(%)

조합	평가지표	SVM	Random Forest	Ada Boost	MLP
8	F1	0.00	22.22	19.05	22.54
	P	0.00	30.00	20.69	21.62
	R	0.00	17.65	17.65	23.53
9	F1	0.00	23.33	21.54	12.50
	P	0.00	26.93	22.59	13.33
	R	0.00	20.59	20.59	11.76
10	F1	70.18	69.56	69.70	92.54
	P	86.96	68.57	71.88	93.94
	R	58.82	70.59	67.65	91.18
11	F1	62.96	71.43	72.73	92.54
	P	85.00	69.44	75.00	93.94
	R	50.00	73.53	70.59	91.18

전체 4개 조합 중 모델의 성능에 유의미한 변화가 있는 부분을 기준으로 조합 8-9, 조합 10-11의 두 개의 그룹으로 나눌 수 있다.

<표 7>을 보면 F2(댐수위차)가 조합에 포함된 경우와 그렇지 않은 경우로 구분할 수 있다. 분석결과 F2(댐수위차)와 F6(댐수위차평균)를 모두 자질 조합에 포함하는 경우 좋은 성능을 나타낸다는 것을 알 수 있다. 2차 실험에서도 MLP가 가장 높은 성능을 보였다.

3-3. 기존 자질의 변경 및 평가

이번 실험에서는 기존 실험에 사용한 F3(댐평균수위)를 그대로 사용하지 않고 F1(댐수위)과의 차이 값을 적용하였다(F3′= F1-F3). 이렇게 하면 F1, F4, F5를 제외하고 F2, F3′, F6 간의 값의 차이를 크지 않게 하는 효과가 있다. 실험 시 조합 11의 F3을 F3′으로 대체하였다.

결과는 <표 9>와 같으며, 비교하기 쉽도록 조합 11의 평가 결과와 같이 표시하였다.

<표 9> 기존 자질의 변경에 따른 평가 결과(%)

조합	평가지표	SVM	Random Forest	Ada Boost	MLP
11	F1	62.96	71.43	72.73	92.54
	P	85.00	69.44	75.00	93.94
	R	50.00	73.53	70.59	91.18
12	**F1**	**64.15**	**76.22**	**76.92**	**96.97**
	P	**89.47**	**80.75**	**80.64**	**100.00**
	R	**50.00**	**73.53**	**73.53**	**94.11**

분석 결과 전체 모델의 정밀도가 모두 상승하였다. 재현율의 경우 Ada Boost와 MLP에서 약 3%포인트 정도 성능이 향상되었으나, SVM과 Random Forest의 재현율에는 변화가 없었다.

3-4. 평가 결과 분석

SVM을 제외하고 나머지 모델은 자질 조합 12에서 가장 높은 성능을 나타냈으며, Random Forest와 Ada Boost는 비슷한 성능을 나타냈다. 전체 모델 중 MLP가 가장 높은 성능을 보였으며, 기존 자질을 활용하여 생성한 자질을 추가하거나 자질 간 차이를 작게 할 때 성능이 향상되는 것을 확인할 수 있었다.[3]

4. 결론 및 향후 연구

이 실험을 통해서 기계학습을 활용하여 효과적으로 댐 수위 이상 데이터를 탐지할 수 있다는 것과 기계학습 모델의 성능은 자질의 수뿐 아니라 적절한 자질의 처리도 영향을 미친다는 것을 알 수 있었다.

이 실험은 기계학습에 기반하여 댐 수위 이상 데이터를 탐지하는 첫 시도라는 점에 큰 의의가 있다.

또한, 댐 수위 이상 데이터의 검출뿐 아니라 정수장이나 가압장과 같은 수도 시설에서 펌프 가동상태의 이상 유무를 탐지하여 시설 침수와 같은 사고 예방 등 그 활용도가 높을 것으로 판단된다.

향후 연구에서는 이번 연구 결과를 바탕으로 이상 데이터 탐지성능에 영향을 미치는 데이터 전처리 방법 및 다른 자질에 관한 연구와 다른 댐에서는 어떤 결과가 나타날지 등을 중심으로 연구를 확대해 보고자 한다.

참고문헌

[1] 건설교통부, “수문 관측 매뉴얼”, p1-2, 2004.
[2] 조우쯔화, “단단한 머신러닝”, 제이펍, p39-40, 2020.
[3] 이항석, “빅데이터 분석에 의한 요율 산정 방법 비교”, 보험연구원, p61-74, 2018.