엔트로피 이론을 이용한 상수처리 약품주입 간이룰의 유도

Derivation of Simplified Rule for Coagulant Dosing in Water Treatment Process using Entropy Minimization

김창종 (수원대)　　백승욱 (숭실대)*

C. J. Kim (University of Suwon)　　Seung Uk Baek(University of Soongsil)

ABSTRACT

정수처리 공정은 응집, 침전, 여과, 살균소독 처리로 이루어지며, 이 가운데 약품주입에 의한 응집 침전 처리는 상수처리시스템에서 가장 핵심 부분이 된다. 투입해야할 약품의 양은 여러 가지 조건에 따라 달라지는데 원수의 탁도, 수온, 그리고 pH등의 여러 요인이 있다. 본 논문에서는 jar테스트의 결과를 이용하여 최초 물의 상태에 따라 약품의 주입량을 간단하게 결정짓는 룰의 유도를 다루고 있다. 이러한 간이룰의 유도를 위하여 엔트로피 이론을 적용하였다. 이 이론을 이용하여 출력변수인 유출수질(탁도)의 만족/불만족의 상태에 따라 입력변수인 유입수질(탁도, 수온, pH, 알칼리도 등)과 결정변수인 약품투입량에 대한 간이룰을 유도하고 시험용 데이터에 이 룰에 적용시켜 그 결과를 알아보았다.

I. 서 론

상수원의 물을 안전한 식수 및 공업용수로 공급하기 위한 시스템을 상수처리시스템이라한다. 상수처리 플랜트의 역할은 양질의 안전한 식수 및 공업용수를 공급하는 것이다[1]. 이러한 목적을 위한 원수의 근원은 지하수, 강, 또는 호수의 표면수 등인데 이 표면수는 다량의 불순물을 포함하고 있어 수질 변화 폭이 크다고 하겠다. 응집제는 이러한 불순물을 완전히 또는 효과적으로 처리하기 위한 화학약품으로, 응집제 주입율은 정수장에 흘러들어오는 원수의 수질변화에 따라 변하게 된다. 상수처리시스템의 정수처리공정은 응집, 침전, 여과, 살균소독 처리 과정을 거치며, 이 가운데 약품주입에 의한 응집·침전 및 살균소독 처리는 상수처리시스템에서 가장 핵심 부분이 된다. 응집제로는 보통 PAC(polymerized aluminum chloride)이 사용되는데 이 응집제 주입은 불순물에 상당히 복잡한 방법으로 영향을 미치기 때문에 공정 모델링이 상당히 까다롭다. 이러한 이유로 주입률을 고정시키는 널리 알려진 방법이 없다[2]. 일반적으로, 상수처리에서 응집과 침전 반응을 고려할 때, 6개의 독립변수 원수 탁도, 수온, pH, 알칼리도, 응집제 주입율, 처리수 탁도가 응집제 프로세스에 상대적으로 관련 정도가 큰 것으로 알려져 있다. 기존의 응집제 주입은 JAR 테스트에 의한 Look-up Table[3]의 이용 또는 통계 팩키지에 의한 선형 모델링 기법[4]에 의하고 있다. 본 논문에서는 Jar 테스트에 의한 응집제 투입의 결과데이터를 이용하여 최초의 수질에 따라 응집제 투입을 엔트로피 이론에 의하여 간단하게 결정하는 방법을 고안하여 응집제 투입 간이룰을 형성하였다.

II. 엔트로피 이론과 클래스 구분

A. 엔트로피 이론

열역학에서 사용되던 엔트로피(Entropy)이론을 Shannon이 최대 통신량의 문제에 응용한바 있다. 이것이 정보이론에 사용되는 엔트로피의 개념이다.[5] 정보이론에서 엔트로피 증가와 감소를 계산하는 방법은 현재 가능한 데이터와 기대치를 비교하는 것이다. 어떤 현상의 결과에 대해 예측이나 예상이 근접할수록, 실제 결과에서 얻을 수 있는 정보의 양은 준다. 즉, 어떤 사건이 이미 예측하는데로 결론지어진다면, 이런 사건에서 얻는 정보의 양은 적음을 알 수 있다. 다시말해, 예상치 못했던 일이 많이 일어날수록 얻을 수 있는 지식은 증가하는 것과 같다. 여기에서, 정보의 증가가 최소가 될 때, 어떤 현상에 대한 예측

을 최적으로 할 수 있다는 것을 알 수 있다. 이를 두 클래스의 구분(Classification)에 적용한다면, 정보의 양이 최소가 되는 점이 두 클래스를 최적으로 분할한다는 것을 알 수 있다. 이 정보의 양(Quantity of information)은 확률치의 로그값에 음(-)의 부호를 붙인 것으로 정의된다.[6]
우선, $P(X_i)$를 i번째 데이터 X_i가 TRUE일 확률이라 가정한다. 새로운 X_i가 TRUE라면 정보의 이득은 $S(X_i) = -KlnP(X_i)$이고, FALSE라면 $S(-X_i) = -Kln(1-P(X_i))$ 의 정보의 이득을 볼 수 있다.
여기에서, 엔트로피를 정보의 기대치로서 정의하고, X_i에서 얻을 수 있는 정보의 기대치를 다음 식으로 구할 수 있다.

$$S(X_i, -X_i) = -K[P_i \ln P_i + (1-P_i)\ln(1-P_i)] \quad (1)$$

따라서, 모든 데이타에 엔트로피는

$$S = -k\sum_{i=1}^{N}[P_i \ln P_i + (1-P_i)\ln(1-P_i)] \quad (2)$$

B. 엔트로피 최소화에 의한 임계치의 계산
그림1에서와 같이 두 종류 샘플의 클러스터링 POINT X를 클래스 사이의 임계치라 한다. 두 개의 클래스를 가지고 있는 Xmin에서 Xmax까지의 샘플에서 두 개의 클래스를 정확하게 나눌 수 있는 최적의 임계치를 찾으려할 때, 임의의 임계치 즉, Xmin과 Xmax사이의 X를 설정하여 아래와 같은 엔트로피의 계산과정을 통해 임계치를 산출할 수 있다[7]. X는 모든 샘플을 p지역[Xmin, X]과 q지역[X, Xmax]으로 나누며 $S(X) = p(X)S_p(X) + q(X)S_q(X)$가 X에서의 엔트로피값이다.

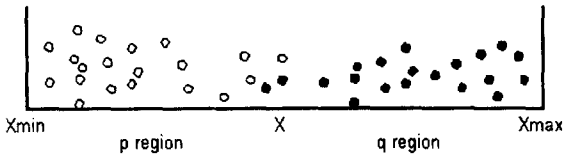

그림 1. 전체 샘플의 분포와 임의의 임계치 X

여기서, p(X)는 p지역에 모든 샘플이 있을 수 있는 확률이며, q(X)는 q지역에 모든 샘플이 있을 수 있는 확률인데 p(X) + q(X) = 1의 관계를 갖고 있다. $S_p(X)$와 $S_q(X)$는 p지역과 q지역의 엔트로피인데 아래와 같이 구해진다.

$$S_p(X) = -(p_1(X)\ln p_1(X) + p_2(X)\ln p_2(X)) \quad (3)$$

$$S_q(X) = -(q_1(X)\ln q_1(X) + q_2(X)\ln q_2(X)) \quad (4)$$

여기에서, $p_k(X)$: p지역에 CLASS k 샘플이 있을 수 있는 확률이며,
$q_k(X)$: q지역에 CLASS k 샘플이 있을 수 있는 확률로 각각을 아래에 나타내었다.

$$p_k(X) = \frac{n_k(X) + w(X)}{n(X) + w(X)} \quad (5) \qquad p(X) = \frac{n(X) + v(X)}{n + v(X)} \quad (6)$$

여기에서, $n_k(X)$는 p지역에 CLASS k 샘플의 수, n(X)는 q지역의 샘플 수, 그리고, n은 p지역과 q지역의 전체 샘플 수를 나타내며, w와 v는 선험적 가중치인데 w와 v를 1로 놓고 간략화하여 쓴다.
$q_k(X)$와 q(X)에 대하여는 위와 같은 방식을 이용하여 구한다. 위와 같은 과정을 반복 최소의 임계치를 가진 X 즉, Xmin과 Xmax사이의 최적의 임계치는 식(7)과 같다.

$$X = Smin(Xmin, Xmax) \quad (7)$$

D. ANALOG SCREENING에 의한 간이룰의 생성
입력 아날로그 변수들로부터 임계치를 기준으로 임계치보다 작으면 0, 크면 1로 변환하여 바이너리 테이블을 만든다. 이 바이너리 테이블을 이용하여 룰을 만든다.[8] 룰은 Decision Tree의 형태를 가진다. 일반적으로 룰을 유도할 때에는 여러 가능한 룰 중에서 최적의 룰을 선택하게 되는데, 이때에도 엔트로피 이론을 적용한다. 룰의 유도를 위한 엔트로피 식은 아래와 같다.

$$S = -k\sum Y_i \cdot P_i \cdot \ln P_i, \qquad i = 1, \ldots\ldots, m \quad (8)$$

여기에서, m은 Decision Tree에서 나누어지는 Branch의 수를 나타낸다.
Y_i는 룰의 i번째의 branch에 의하여 처리되는 모든 샘플의 수를 나타낸다. 즉, i번째의 branch에서 두 클래스로 나눌 때 에러 없이 나누어 지게되는 샘플의 수.
P_i는 branch i에서 두 클래스로 나눌 때, 각 클래스로 나누어지는 샘플에 대한 평균확률(mean probability)을 나타낸다. k는 상수로서 최소 엔트로피를 구하는데 있어서는 그 역할이 없다.

위의 엔트로피 식을 사용하여 모든 branch i에서의 모든 경우에 대한 Y와 P를 계산하여 엔트로피가 최소가 되는 Route를 찾으면 그것이 최적의 Rule이 된다. 룰의 평균확률(mean probability 또는 weight)은 $p=(z+t)/(n+t+f)$로 정의된다.[6]
여기에서, t는 구분이 가능한 true(good)상태의 수, f는 구분이 가능한 false(no-good)상태의 수
두 개의 클래스만이 존재할 때 t=1, f=1이 되어 평균확률은 $p=(z+1)/(n+2)$로 간략화된다.

III. 응집제 주입율에 대한 간이률 유도

A. Jar-Test

일반적으로 상수처리 플랜트에서 원수는 화학물질의 투입으로 즉각적으로 정수되어진다. 화학물질의 양은 원수의 수질(원수 탁도, 수온, pH, 알칼리도 등)에 의해 결정되어진다. 즉, 상수처리에서 응집과 침전 반응을 고려하여, 6개의 독립변수(원수 탁도, 수온, pH, 알칼리도, 응집제 주입율, 처리수 탁도)를 입출력변수로하여 최적의 응집제 주입율을 결정한다. 이를 위한 데이터는 Jar-Test를 이용하여 얻었다. 이 방법은 샘플 테스트로 5~6개의 비이커에 원수를 넣어 응집제 주입율을 바꾸어 주입하여 탁질의 침전상태를 보고 이 중에서 탁질의 제거 상태가 제일 좋은 주입율을 선택하는 방법이다. 그리고, 최종 수질이 만족스러울 경우에는 PAC주입이 가장 낮은 것을 최적의 PAC주입량으로 정하는 것이다. 표1에서 (원수)탁도, 수온, pH, Al(알카리도), PAC 그리고 (처리후)탁도를 알 수 있다.

표 1. Jar-Test 데이터

급속교반(100 rpm) (1분간)　완속교반(50 rpm) (5분간)　침수시간 (15분간)

시험일시	94. 4. 7.	시 험 번 호	PAC (ppm)	21 분후			판 정
				탁 도	pH	Al	
채수지점	한강	1	8	2.2	7.5	45	
천 기	맑음	2	10	1.9	7.5	45	
탁 도	1.7	3	12	2.0	7.4	44	
수 온	28.2	4	14	1.6	7.4	44	
pH	7.5	5	16	1.2	7.4	43	○
Al	49	6	18	1.2	7.3	43	

이 방법은 테스트 시간이 오래 걸리나 테스트 시에 바꾸어 넣는 주입율의 선택이 숙련된 Jar 테스트 전문가에 의존하기 때문에 그 결과에 전문가의 전문 지식이 포함되어 있다. 따라서, 응집제 주입공정 간이률 유도에 전문가의 지식이 포함된 Jar 테스트 데이터를 이용하였다.

B. 간이률의 생성

서울의 한 정수장에서 구하여진 Jar테스트의 결과를 이용하여 만족할만한 처리후 탁도가 1.2NTU를 기준으로 Jar 테스트 데이터들을 good, no-good의 두 class로 나누었다. no-good인 경우의 결과는 Rule 생성에 의미가 없으므로 good인 경우의 샘플 데이터만을 이용하여 룰을 생성한다. good 인 경우를 class1이라하고 각각의 변수들을 A, B, C, D라 한다. class1인 샘플들에서 엔트로피 최소화 법칙을 이용하여 A, B, C, D변수의 에너지최소점인 threshold 값을 구한다. 이를 기준으로 각각의 변수들을 바이너리 값으로 변환한다. 이 바이너리 값들을 가지고 Analog screening을 통해 생성한 semi-rule은 If A is 0 then class 1. weight=0.9375이며, 이것을 이용하여 다음과 같이 rule을 만들 수 있다. 즉, "A가 0인 경우의 data들이 가지고 있는 PAC 주입량들 중에서 가장 최소가 되는 양을 PACs라고 정한다면, If A is 1 then PAC is PACs라는 rule이 결정된다. 또한, 이 rule의 신뢰도는 0.9375이다."

C. PAC 주입율의 결정률 생성

정수처리 시스템의 엔트로피 최소화 규칙적용을 통해 얻은 semi-rule은 If A is 0 then class1 0.9375이며, 이 semi-rule을 이용하여 최종 결과를 아래와 같이 생성할 수 있다. 여기에서, A는 온도, 원수탁도, pH, 알카리도 등을 나타내며, 이것이 '0'라고 하면 A의 임계치 값보다 작은 값들의 집합을 의미한다. PACs란 A가 0인 경우와 대응하는 PAC 주입량 중에서 최소가 되는 양을 뜻한다. 즉,

$$PACs=\min\{A\{PAC\}\} \tag{9}$$

여기에서, A{PAC}란 A에 대응하는 PAC 주입량의 집합을 나타낸다.
본 논문에서는 룰을 생성하기 위해, jar 테스트 결과(323개)에서 무작위로 데이터를 추출하여 학습용 데

이터(160개)와 테스트용 데이터(160개)로 나누었다.
위와 같은 과정을 거쳐, jar-test의 데이터 중 학습용 데이터를 이용하여 최종룰을 생성하였다. 즉,

If A=원수탁도 is 0 then PACs=min{A(PAC)}=17 0.9375 (9)
else If B=온도 is 0 then PACs=min{B(PAC)}=19 0.8353 (10)
else If D=알카리도 is 0 then PACs=min{D(PAC)}=20 0.7500 (11)
else If C=pH is 1 then PACs=min{C(PAC)}=21 0.7167 (12)
endif

이것의 의미를 다시 쓰면 아래와 같다. 즉, 최종 PAC주입량 룰은
" 원수의 탁도가 17.26NTU보다 작다면, PAC투입량은 17ppm이다. 이 룰의 신뢰도는 0.9375
아니면 원수의 온도가 20.095℃보다 작다면, PAC 투입량은 19ppm이다. 이 룰의 신뢰도는 0.8353
아니면 원수의 알카리도가 31.5보다 작다면, PAC 투입량은 20ppm이다. 이 룰의 신뢰도는 0.7500
아니면 원수의 pH가 7.04보다 크다면, PAC 투입량은 21ppm이다. 이 룰의 신뢰도는 0.7167 " 이다.

D. PAC주입량 결정룰의 테스트
학습에 의한 결과는 (9),(10),(11),(12)와 같이 나왔다. 이를 테스트 하기 위해 테스트용 데이터를 이용하여 최종룰을 생성한 결과, 아래와 같이 나왔다.

If A=원수탁도 is 0 then PACs=min{A(PAC)}=17 0.9565 (13)
else If D=알카리도 is 0 then PACs=min{D(PAC)}=20 0.7826 (14)
else If B=온도 is 0 then PACs=min{B(PAC)}=19 0.7609 (15)
else If C=pH is 1 then PACs=min{C(PAC)}=21 0.7500 (16)
endif

D(알카리도)와 B(온도)의 순서가 바뀌고 중요도가 좀 달라졌다. 또한, 중요도가 약간 달라졌을 뿐 PAC주입량의 변화는 없었다. 이를 보아 학습용 데이터의 결과를 신뢰할 기반이 생겼다.
테스트용 데이터를 학습용 데이터를 이용하여 얻은 최종룰에 적용하였다. 그때, 나온 결과를 기존의 jar테스트 결과와 비교하여 본 결과, 더욱 낮은 PAC주입을 한다는 것을 알 수 있다. 이것이 엔트로피 이론의 장점이다. 하지만, 충분한 데이터의 검증은 이루어지지 않았다. 더욱 많은 데이터의 적용과 실 플랜트의 적용문제가 남아 있다.

IV 결론

정수장에서의 정수처리 공정에 가장 중요한 응집제 주입량을 jar테스트로부터 결정하기 위하여 엔트로피 이론을 적용하였고, 이 이론에 따라 응집제에 대한 룰을 생성하였다. 이 룰은 최종 수질이 좋다면 가장 최소의 PAC을 주입하면 된다는 근거에 의하여 PAC 주입율에 대한 상수처리 간 이룰을 생성하여 이 룰을 실험 데이터에 적용하여 그 적용 가능성을 보였다.

V. 참고 문헌

[1] Clark, J.W., W. Viesman, Jr. and M. J. Hammer, "Water Supply and Pollution Control", International Textbook Co., 1971.
[2] I. Enbutsh, K. Baba and N. Hara, " Fuzzy rule extraction form a multilayered neural network," Proceedings of IJCNN '91, Seattle, 1991.
[3] Lin-X. Wang, Adaptive Fuzzy Systems and Control: Design and Stability Analysis, Prentice-Hall, Englewood Cliffs : NJ, 1994.
[4] S. Watanabe, et. Al., "Intelligent Operation Support Systems for the Activated Sludge Process," Wat. Sci. Tech., Vol. 28, 1993
[5] R. Christensen, Fundamentals of Inductive Reasoning, Entropy Ltd., Lincoln, MA, 1980.
[6] R. Christensen, Entropy Minimzax Sourcebook, vol. I-IV, Entropy Ltd., Lincoln, MA, 1980.
[7] C. J. Kim, "An Intelligent Decision-Making System for Detecting High Impedance Faults," Ph.D. Dissertation, Texas A&M University, College Station, Texas, December 1989.
[8] C. J. Kim, B. D. Russell, "Classification of Faults and Switching Events by Inductive Reasoning and Expert System Methodology," IEEE Transactions on Power Delivery, vol. 4, no.3, pp. 1631-1637, July 1989.