Journal of the Korean Society of Water and Wastewater
Vol. 20, No. 6, pp. 789-797, 2006

상하수도학회지, 논문
20권 6호, pp. 789-797, 2006

Model Tree기법을 이용한 정수처리공정에서의 응집/침전 효율 예측에 관한 연구

Establishment of the Refined Model for Prediction of Flocculation/Sedimentation Efficiency Using Model Tree Technique

박노석* · 박상영 · 김성수 · 정남정 · 이선주

No-Suk Park* · Sang-Young Park · Seong-Su Kim · Nam-Jeong Jeong · Sun-Ju Lee

한국수자원공사 수자원연구원

(2006년 6월 16일 논문 접수; 2006년 12월 11일 최종 수정논문 채택)

Abstract

This study was conducted to establish the refined model for prediction of flocculation/sedimentation efficiency in factual drinking water treatment plants using model tree technique. In order to carry out machine leaning for determining each linear model, five parameters; time, coagulant dose, raw water turbidity, SCD and conductivity, which were measured and collected from the field (K_DWTP), were selected and used. The existing analytical models developed by previous researchers were used only to examine closely the mechanism of flocculation rather than to apply it for practical purpose. The refined model established using model tree technique in this study could predict the factual sedimentation efficiency accurately (below 9% of average absolute error). Also, in aspect of engineering convenience, without any additional manipulation of parameters, it can be applied to practical works.

Key words: model tree technique, prediction of flocculation/sedimentation efficiency, the refined model
주제어: 모델트리기법, 응집/침전 효율 예측, 개선 모델

1. 서 론

응집은 침전속도가 낮은 콜로이드 입자의 제거에 범용적으로 사용되는 기작이다. 왜냐하면 자연계에서 생성되는 대부분의 콜로이드 입자는 음(negative)으로 대전(charge)되어 입자간의 반발력에 의해 안정된 부유상태로 존재한다(Letterman et al., 1999; Reynolds and Richards, 1996).

이에 금속염계 응집제가 물에 주입되면, 용해되어, 금속 이온이 수화되어 양(positive)으로 대전된 수화금속계 착화합물을 생성한다. 이러한 수화금속계 착화합물(hydroxometallic complexes)이 콜로이드입자의 표면에 흡착되면, 전기적인 반발력이 콜로이드 입자

*Corresponding author Tel: +82-42-860-0390, FAX: +82-42-860-0399, E-mail: nspark@kwater.or.kr (Park, N.S.)

가 불안정화될 정도로 감소된다. 불안정화된 입자는 침전과 여과에서 제거 가능한 플록 생성을 위한 응집 과정에서 다른 입자들과 충돌에 의해 성장하게 된다. 이러한 응집기작을 "전하 중화(charge neutralization)"이라 한다. 물속에서 금속계 착화합물의 생성은 불용해성의 금속 수산화물의 생성과 동시에 수반된다. 수산화물의 생성은 응집제의 특성 및 주입량에 따라 달라지며, 생성된 금속 수산화물(metal hydroxide)은 불안정화된 콜로이드 입자의 표면에 흡착되어 "enmesh"를 형성, 입자를 침전시킨다. 이러한 기작을 "sweep coagulation"이라 한다(Letterman et al., 1999; Reynolds and Richards, 1996).

이상적으로, 응집제 주입량은 효율적인 플록의 생성을 위해 불안정화가 일어날 수 있도록 충분하여야 한다. 실제, 일반적으로 안정성을 고려하여 응집제는 약간 과도하게 주입하고 있다. 이러한 과도한 응집제의 주입은 수화금속계 착화합물의 부가적인 생성을 일으켜, 이미 불안정화된 플록 표면의 전하를 다시 양으로 하전되게 하여 안정화되는 결과를 가져올 수 있다. 반면에 과도한 응집제의 주입은 금속 수산화물의 형성을 증가시켜, 재안정화된 입자의 표면에 흡착되어 sweep coagulation의 기작을 증대시킬 수 있다. 이에 어느 정도의 입자의 재안정화 현상을 상쇄시킬 수 있다.

유동 전류계(SCD; Streaming Current Detector)는 응집제 주입후 샘플링한 물 내의 입자 표면의 상대적인 전위를 측정하는 장치이다. SCD는 1960년대 중반에 하수처리에 첨단 기술로 소개된 이후 미국 등지에서 on-line monitoring 및 응집제 자동 주입 제어에 범용적으로 사용되는 기술이 되었다(Walker et al., 1996).

다음 Fig. 1은 간략하게 SCD의 작동 원리를 설명하고 있다. 그림에서 보는 바와 같이 하전된 입자를 가진 물이 피스톤에 의해 원통형 실린더에 들어오고 나간다. 이 과정에서 피스톤의 움직임과 내부 전극(electrodes)에 의해 교류(alternating current)가 발생하는데, 이 전자의 흐름을 유동전류(streaming current)라 한다. 이 유동전류는 입자에 대전된 전하의 양에 비례한다. 기존 연구에 있어서도 SCD output과 zeta potential과는 밀접한 상호 관련성이 있는 것으로 밝혀졌다(수자원공사, 2004).

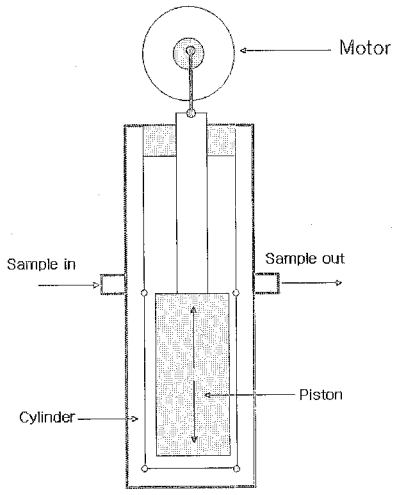


Fig. 1. SCD 작동 원리.

SCD는 "feed back" 형식의 응집제 제어를 한다. 즉 응집제가 주입된 후 물을 샘플링하여 분석한다. 샘플링한 물이 SCD 센서를 통과하면서 발생된 유동전류(streaming current)는 증폭되어 전압(mV)의 단위로 표시되어 진다. 이 값은 사전에 결정된 "set point"를 근거로한 응집제 주입량의 결정에 이용된다(Dentel and Kingery, 1988).

상기 언급한 SCD는 feed back 제어를 근간으로 하기 때문에 혼화공정에서 응집제가 과다 및 과소로 주입된 원수는 대처하기가 어렵다. 즉, 이미 응집제가 주입된 원수는 SCD에 의해 지속적으로 측정하는 것은 가능하지만, 적절한 응집조건에 의해 발생하는 응집효율을 예측하는 기능은 미비한 것으로 판단된다.

한편 이제까지 정수처리공정에서 발생하는 floc의 응집 및 침전의 효율을 예측하는 실질적인 예측 모델은 전무한 실정이다. 단지 Smoluchowski(1917)가 제안한 층류장에서 두 입자의 충돌주기함수를 수학적으로 표현하는 모델이 있었으며, 이후 Tambo와 Watanabe(1979)는 난류장에서 입자간의 충돌을 통계학적으로 표현하는 해석적 모델을 제시한 바 있다. 이후 Han과 Lawler(1992)는 이러한 크기가 다른 입자들간의 충돌에 영향을 미치는 힘(force)를 수식을 이용하여 정량화하려는 연구도 수행한 바 있다. 그러나 이러한 모델은 실제 공정에서의 응집 및 침전 효율을 예측하고자하는 목적보다는 메커니즘을 규명하기위해 수립된 해석 모델이라 실공정상에서 발생하는

floc의 응집 및 침전 효율을 예측하기에는 무리가 있다.

이에 본 연구에서는 실제 정수처리공정에서 응집공정의 효율에 영향을 미치는 실제적인 인자(실시간으로 변화 측정이 가능한 인자)를 도출하였으며, 이러한 인자들을 model tree기법을 통해 기계학습(machine learning)시켜 floc의 침전효율을 예측하는 모델을 수립하고자 하였다.

2. Decision Tree 및 Model Tree 기법

Decision Tree는 분석용 자료를 이용하여 의사결정 규칙을 나무구조로 만들고, 관심대상을 몇 개의 하위 집단으로 분류하거나 예측을 하는 데이터 마이닝(data mining) 기법이다. 이러한 decision tree는 분석과정이 나무구조에 의해서 표현되기 때문에, 분류 또는 예측을 목적으로 하는 다른 데이터 마이닝 기법들에 비해, 분석과정의 이해와 결과의 해석이 쉽다는 장점을 가지고 있다.

특히, 데이터 마이닝에서 decision tree의 활용이 많은데, 이는 decision tree 자체가 분류 또는 예측 모형으로 활용되어 데이터 마이닝 기법으로 사용되기도 하고, 다른 데이터 마이닝 기법을 적용하기 전에 데이터를 처리할 수 있는 작업에도 사용할 수 있기 때문이다. Decision tree 기법은 연속형이나 범주형 등의 예측변수를 그대로 이용하므로 데이터의 변형에 필요한 시간을 줄일 수 있고, 모형을 구축하는 시간이 짧다. 이러한 특성 때문에 Decision tree 기법은 다른 예측 기법을 수행하기 전에 많은 예측변수 중에서 유용한 것들만 고르는 과정에 사용될 수 있다.

Decision tree는 뿌리마다 시작해서 각 가지가 끝마디가 될 때까지 자식마디를 만들면서 형성된다. 이러한 decision tree를 완성하기 위해서는 마디의 분리기준(splitting rule)의 선택, 분리를 멈추기 위한 정지기준(stopping rule)의 선택, 가지치기(pruning)방법의 선택, 입력변수의 값에 결과치가 있는 경우 결측치 대치(imputation) 방법의 선택 등 여러 단계를 수행하여야 한다. 이러한 과정을 수행하여 decision tree를 형성하는 주요 알고리즘으로는 CHAID, CART, C4.5, QUEST 등이 있으며, 각 단계에서 서로 다른 기준을 가지고 있어 다른 decision tree가 만들어진다(최종후 et al., 2003). Table 1은 몇 가지 관점에서 이들 알고리즘을 비교한 것이다.

본 연구에서는 C4.5 알고리즘을 개량하여 기계학습 패키지 중 하나인 WEKA(Witten and Frank, 1999)로 구현한 M' 5(prime) 알고리즘을 사용하여 연구를 수행하였다. M' 5알고리즘은 세 개의 모델 즉, 선형회귀식, 회귀나무(regression tree), 모델 트리(model tree) 분석을 수행한다. 선형회귀식은 n개의 서로 다른 입력 변수에 대한 회귀식을 제공한다. 회귀나무와 모델트리는 의사결정나무의 결과로서 수치형 결과를 제공하는 면에서는 유사하나, 회귀나무 알고리즘은 출력 결과에 대한 평균값을 제시하고 모델트리는 선형회귀식을 제공한다는 면에서 차이가 있다.

모델트리는 의사결정나무 알고리즘의 잎노드를 선형 회귀함수로 제시한다. 따라서 모델트리는 의사결정구조를 명료하게 제시함과 동시에 결과로써 제시된 선형함수는 일반적으로 많은 변수를 포함하지 않는다는 장점을 가지고 있다. M' 5 모델트리는 부분적으로 선형모델로써 선형모델인 ARIMA와 비선형모델인 인공신경망(ANN)의 중간정도 위치를 갖는다. 분류문제에 있어서 의사결정나무 알고리즘은 다음과 같은

Table 1. Comparison of the Decision Tree Algorithm

구분	CHAID	CART	QUEST	C4.5
목표변수	명목형, 순서형, 연속형	명목형, 순서형, 연속형	명목형	명목형, 순서형, 연속형
예측변수	명목형, 순서형, 연속형 (사전 그룹화)	명목형, 순서형, 연속형	명목형, 순서형, 연속형	명목형, 순서형, 연속형
분리기준	카이제곱-검정 F-검정	지니 지수 분산의 감소	카이제곱-검정 F-검정	지니 지수 분산의 감소
분리개수	다지분리(multiway)	이지분리(binary)	이지분리(binary)	다지분리(multiway)
가지치기	○	○	○	○
결손 값 대체	×	○	○	○
비용함수	×	○	○	○

분할 정복(divide-and-conquer) 방법을 사용한다.

(1) 뿌리마디에 위치할 변수를 선택하고 가능한 값에 대하여 하나의 가지를 생성한다.

(2) 선택된 값에 대하여 자식마디를 분리 생성한다.

(3) 자식마디에 할당된 모든 샘플이 하나의 범주로 분리될 때 자식마디의 분리 생성을 중지한다.

모델트리에서 통계학적 분리기준은 자식마디의 엔트로피(entropy, 불순도(不純度)) 감소에 근거하고 있다. 즉 유사한 서질의 샘플을 가능한 하나의 자식마디로 분류하는 방식으로 분리를 수행한다. M' 5모델트리는 자식마디 T에 할당된 샘플의 표준편차가 줄어드는 방향으로 분리를 수행한다. 각 마디에 할당된 자료의 표분 편차는 예측오차로써 평가되고 표준편차의 감소량을 최대화 시키는 변수가 분리기준으로 선정된다.

$$SDR = sd(T) - \sum \frac{|T_1|}{|T|} \times sd(T_i) \qquad 식(1)$$

여기서, SDR(Standard deviation reduction)은 표준편차의 감소량, T_i는 선정된 변수에 의해 생성된 자식마디의 샘플 집합이다. 나무구조의 분리는 표준편차 변화가 미미하거나(약 5% 미만) 자식마디에 할당된 샘플 수가 거의 없을 때 중지된다. 최종적으로 각각의 자식마디 샘플에 대하여 선형회귀모델을 구축한다.

3. 적용 변수 및 데이터

Model tree기법을 이용하여 응집효율을 예측할 모델을 만들기 위해서 본 연구에서는 실제 K_정수장을 대상으로 혼화공정을 전후로 실시간으로 측정하고 있는 인자들을 선정하였다. 그 인자들은 시간(min), 응집제(PAC) 주입량(mg/L), 원수의 탁도(NTU), SCD값(mV), conductivity(mV)이다. 이러한 인자들을 독립변수로 학습시켜 예측할 대상 변수는 floc들의 침강이 일어난 침전공정에서 유출된 처리수의 탁도(NTU)이다.

다음 Fig. 2는 처리수 탁도에 영향을 미치는 각각의 인자들과 시간과의 상관성을 나타내는 그래들이다.

Fig. 2에서 나타나듯이 각각의 변수들은 시간에 대해 일정한 경향성을 보이지 않고 있다. (a) 응집제 주입량과 (b) 원수의 탁도 간에 다소 상관성을 보이기는 하지만 scale상으로 판단하건데 밀접한 상관성을 보인다고 결론내리기에는 무리가 있다.

이에 본 연구에서는 (a) 응집제 주입량, (b) 원수 탁도, (d) SCD값(서론에서도 언급하였듯이 Zeta potential과 비슷한 경향을 보임) 마지막으로 (e) 전기전도도 등의 데이터를 이용하여 예측변수인 (c) 처리수의 탁도를 예측하는 모델을 수립하고자 한다.

4. 모델 수립 및 검증

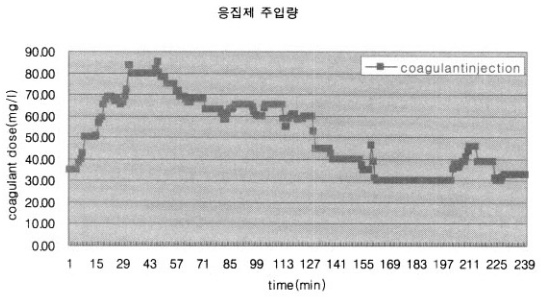
본 연구에서는 실 공정에서의 응집효율을 예측할 수 있는 모델을 수립하기 위해 실제 K_정수장에서 측정되고 있는 실시간 데이터(원수 탁도, 응집제 주입량, SCD값, 전기 전도도)을 이용, Model tree기법을 이용하여 처리수의 탁도를 예측하고자 하였다. 응집효율과 침전의 관계는 이미 기존 문헌연구에서 밝힌바, 반드시 floc의 크기가 클수록 침전이 잘 일어나는 것이 아니라 floc이 얼마나 조밀하게 응집이 되었는가가 효율 결정에 관건이 된다(Letterman 등, 1999)고 조사되었다. 이에 본 연구에서는 처리수의 탁도를 간접적인 응집효율의 지표로 선정하였다.

본 연구에서는 최고의 이득비율을 얻기 위한 끝가지의 분리 및 최적 회귀식을 만들기 위해 뉴질랜드 University of Waikato에서 개발 보급중인 " WEKA (Waikato Environment for Knowledge Analysis)"를 이용하여 모델을 수립하였다. 이 WEKA는 상기 2장에서 언급한 일련의 과정을 자동으로 수행하여 최적의 예측 모델을 수립하는데 도움을 준다.

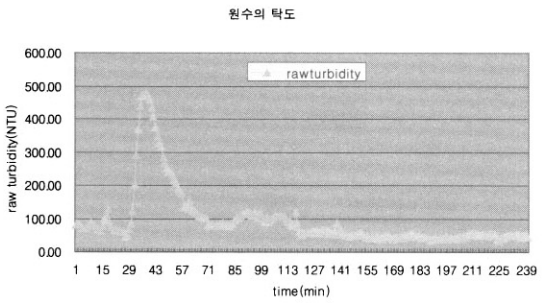
모델을 수립하는데 이용한 데이터는 다음과 같은 회귀형식으로 구성되어 있다.

$$\begin{aligned}[\text{처리수의 탁도(NTU)}] &= a[\text{시간(min)}] \\ &+ b[\text{응집제 주입량(mg/L)}] \\ &+ c[\text{원수의 탁도(NTU)}] \\ &+ d[\text{SCD(mV)}] + e[\text{conductivity(mV)}] \end{aligned} \qquad 식(2)$$

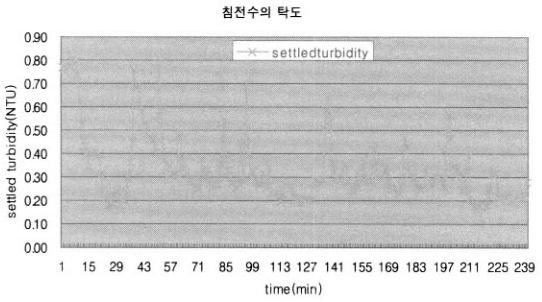
모델 수립을 위해서 상기 데이터는 총 240set을 수립하였으며, 이중 120set은 모델을 수립하는데 사용하였고 나머지 120set은 수립된 모델을 검증하는 데



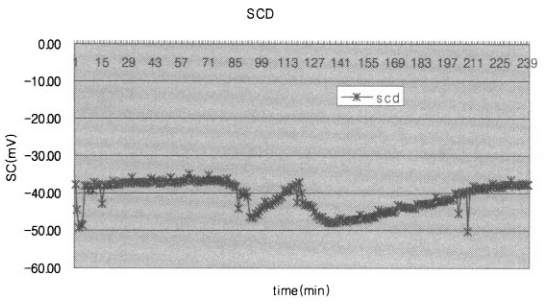
(a) 응집제 주입량 vs. 시간



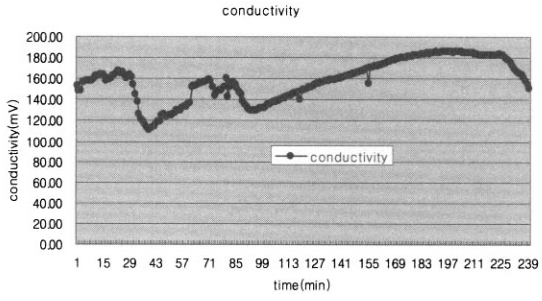
(b) 원수의 탁도 vs. 시간



(c) 처리수의 탁도 vs. 시간



(d) SCD값 vs. 시간



(e) 전기 전도도 vs. 시간

Fig. 2. 적용 변수와 시간과의 상관성 그래프.

에 사용하였다.

4.1. 시간, 응집제 주입량, 원수의 탁도, SCD 및 conductivity를 변수로 사용하는 경우

다음 **Fig. 3**은 상기 언급한 5개의 변수를 모두 회귀식 인자로 사용하여 model tree 예측 모델을 구성한 개요이다. 모두 11개의 끝가지가 구성되어 있으며, 각각의 경우 LM은 "Linear Model"의 약어이며, 각 LM 라벨에 들어 있는 숫자는 회귀식을 구성하는 데이터의 수를 비율로 표시한 것이며, %기호 앞의 숫자는 "relative absolute error"를 표시하고 있다.

다음 **Table 2**는 상기 model tree의 각 끝가지(총 11개)의 linear model을 수식으로 정리한 것이다.

Table 2에서 정리한 model tree에 의해서 수립된 모델은 총 240set 중 50%인 120set의 데이터를 이용하여 기계학습을 시켜 각 LM 식에 나타난 변수 앞의 계수값을 결정하였다. 이렇게 수립된 모델을 검증하기 위하여 기계학습에 이용된 이외의 데이터 120set을 이용하여 다음 **Fig. 4**와 같이 예측치와 실측치를 비교하는 그래프를 도시하였다.

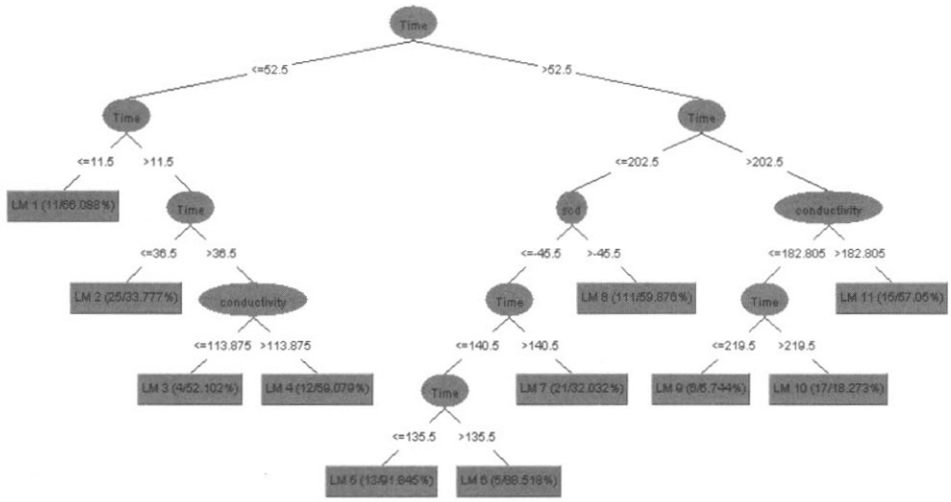


Fig. 3. 시간, 응집제 주입량, 원수의 탁도, SCD 및 conductivity를 변수로 사용하는 경우 Model tree의 구성.

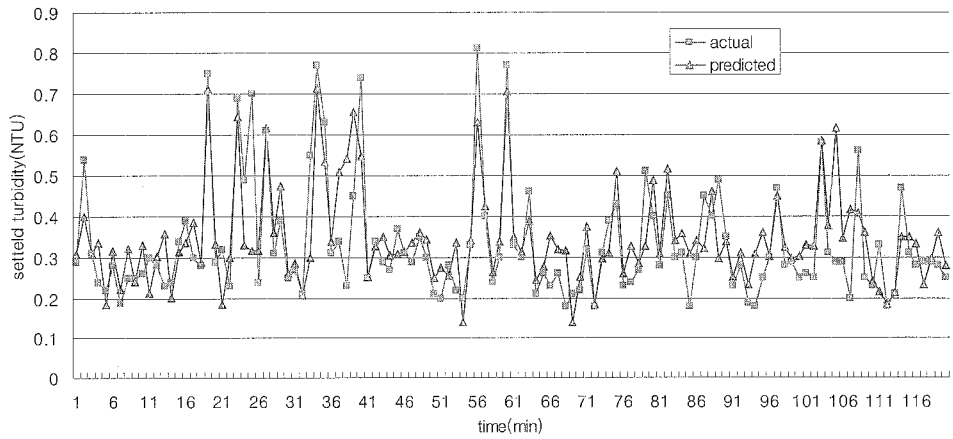


Fig. 4. 시간, 응집제 주입량, 원수의 탁도, SCD 및 conductivity를 변수로 사용하는 경우 Model tree의 검증 결과

상기 Fig. 4에서 제시된 그래프는 Model tree를 이용하여 수립한 모델로 예측한 침전수의 탁도(Table 2에 제시한 식을 이용)와 실제 K_정수장 침전지 유출수의 탁도를 비교한 것이다. 전체적으로 시간에 따라 변동하는 침전수의 탁도 경향을 잘 예측하는 것으로 나타나며, correlation coefficient는 0.7336이고, 평균 error는 약 6.9%로 나타났다.

4.2. 응집제 주입량, 원수의 탁도, SCD 및 conductivity를 변수로 사용하는 경우

상기 4.1에서 제시한 5개의 변수를 이용하여 침전수의 탁도를 예측하는 모델에서 "시간"을 변수로 포함하는 것에 대한 의미에 많은 의문을 가지게 된다. 그 이유는 시간을 제외한 응집제 주입량, 원수의 탁도, SCD값 및 conductivity는 직접적으로 침전수의 탁도에 영향을 주는 인자로 판단이 되지만 시간의 경우 침전수 탁도에 직접적인 영향을 주기보다는 데이터의 수집 시점을 지정하는데 쓰이는 단순한 인자이기 때문이다. 이에 본 절에서는 시간 변수를 제외하고 나머지 4개의 변수를 이용한 model tree LM을 수립하고 이 모델의 예측능을 검증해보고자 하였다. 다음 Fig. 5는 언급한 시간을 제외한 4개의 변수를 회귀식 인자로 사용하여 model tree 예측 모델을 구성한 개요이다.

Table 2. 시간, 응집제 주입량, 원수의 탁도, SCD 및 conductivity를 변수로 사용하는 경우 Model tree의 끝가지 LM

LM No.	예측 변수	LM
1	처리수 탁도(NTU)	= - 0.0085 × 시간(min) - 0.0022 × 응집제 주입량 (mg/L) + 0.0005 × 원수 탁도 (NTU) - 0.0103 × SCD(mV) - 0.0005 × conductivity(mV) + 0.3984
2	처리수 탁도	= - 0.0038 × 시간 - 0.007 × 응집제 주입량 + 0.0008 × 원수 탁도 + 0 × SCD - 0.0005 × conductivity + 0.8839
3	침전수 탁도	= 0.0017 × 시간 - 0.0071 × 응집제 주입량 + 0.0006 × 원수 탁도 + 0.0189 × SCD - 0.0027 × conductivity + 1.831
4	침전수 탁도	= 0.0017 × 시간 - 0.0092 × 응집제 주입량 + 0.0006 × 원수 탁도 + 0.0313 × SCD - 0.0021 × conductivity + 2.3644
5	침전수 탁도	= 0.0013 × 시간 - 0.0021 × 응집제 주입량 + 0.0034 × 원수 탁도 - 0.0119 × SCD + 0.0002 × conductivity - 0.5481
6	침전수 탁도	= 0.0027 × 시간 - 0.0021 × 응집제 주입량 + 0.004 × 원수 탁도 - 0.0119 × SCD + 0.0002 × conductivity - 0.7441
7	침전수 탁도	= 0.0003 × 시간 - 0.0021 × 응집제 주입량 + 0.0011 × 원수 탁도 - 0.011 × SCD + 0.0002 × conductivity - 0.2544
8	침전수 탁도	= -0.0008 × 시간 - 0.0044 × 응집제 주입량 + 0.0012 × 원수 탁도 + 0.0035 × SCD + 0.0009 × conductivity + 0.5362
9	침전수 탁도	= -0.0007 × 시간 - 0.0028 × 응집제 주입량 + 0.0004 × 원수 탁도 + 0.0039 × SCD + 0.0001 * conductivity + 0.5963
10	침전수 탁도	= -0.0009 × Time - 0.0028 × 응집제 주입량 + 0.0004 × 원수 탁도 + 0.0022 × SCD + 0.0001 × conductivity + 0.5668
11	침전수 탁도	= -0.0013 × 시간 - 0.0031 × 응집제 주입량 + 0.0004 × 원수 탁도 - 0.0011 × SCD + 0.0001 × conductivity + 0.5737

Fig. 5에서 제시된바와 같이 총 LM은 10개 이며 각각의 LM은 다음 Table 3과 같다. Table 3 마지막 column에서 예측하는 식이 상수 값으로 제시된 것은 시간이 변수로 제외된 경우 예측능을 극대화하기 위해서 model tree 내부적으로 변형된 형태로 나타난 것이다.

시간 변수를 제외하고 나머지 변수를 이용하여 침전수 탁도를 예측하는 모델을 수립한 경우 상기 표 3과 같이 각 LM이 식으로 나타나지 않고 상수항으로 표현되었다. 이는 Table 2에서 제시한 시간 변수가

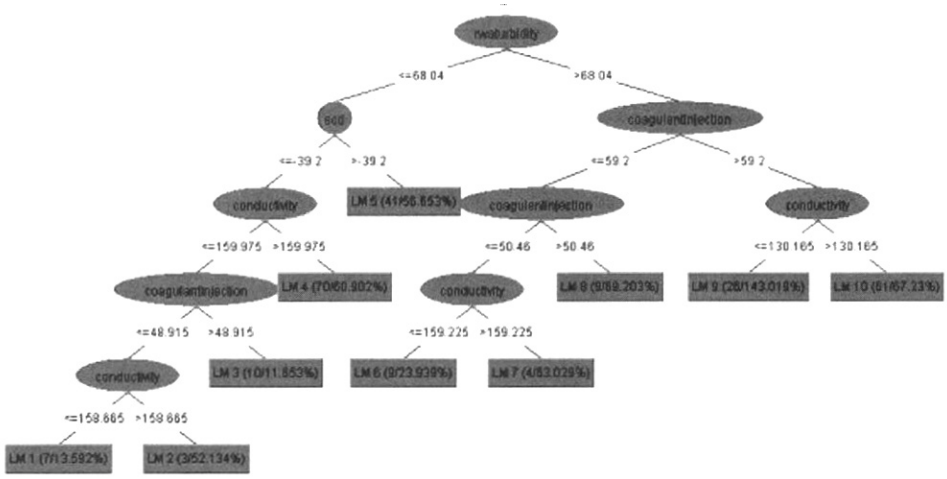


Fig. 5. 응집제 주입량, 원수의 탁도, SCD 및 conductivity를 변수로 사용하는 경우 Model tree의 구성.

Table 3. 응집제 주입량, 원수의 탁도, SCD 및 conductivity를 변수로 사용하는 경우 Model tree의 끝가지 LM

LM No.	예측 변수	LM
1	침전수 탁도(NTU)	= +0.3046
2	침전수 탁도	= +0.3157
3	침전수 탁도	= +0.2807
4	침전수 탁도	= +0.3064
5	침전수 탁도	= +0.2517
6	침전수 탁도	= +0.51
7	침전수 탁도	= +0.4946
8	침전수 탁도	= +0.4182
9	침전수 탁도	= +0.4244
10	침전수 탁도	= +0.3251

없어짐에 따라 연동되어 만들어진 LM이 단순화되어지는 경향을 보이는 것이다. 이에 시간 변수가 포함된 model tree 모델에 비해 예측능이 감소되는 경향을 보였다. 다음 Fig. 6은 Table 3에서 제시한 모델식을 이용하여 모델 수립에 이용하지 않은 데이터 120set을 이용하여 검증한 결과이다.

Fig. 3에서 나타나듯이 침전수의 탁도 변화 경향은 잘 예측하고 있으나, 전체적으로 부분적인 peak치의 절대값은 예측하지 못하는 결과를 보이고 있다. 이에 correlation coefficient는 0.6374, 평균 error값은 8.8% 정도로 나타났다. 또한 그림 6에서 제시한 상기 예측치와 실측치간의 오차를 줄이기 위해 두 값의 최대값으로 나누어 normalization을 한 값들을 비교한 것이기에 Fig. 4에서 제시한 검증 데이터와 절대값이 차이가 발생한 것이다. Fig. 6의 데이터 값을 normalization시킨 것은 각 변수들이 가지는 값들의 order수를 통일시키기 위함이다. 예를 들어 conductivity의 범위는 10~100mV를 가지지만 원수

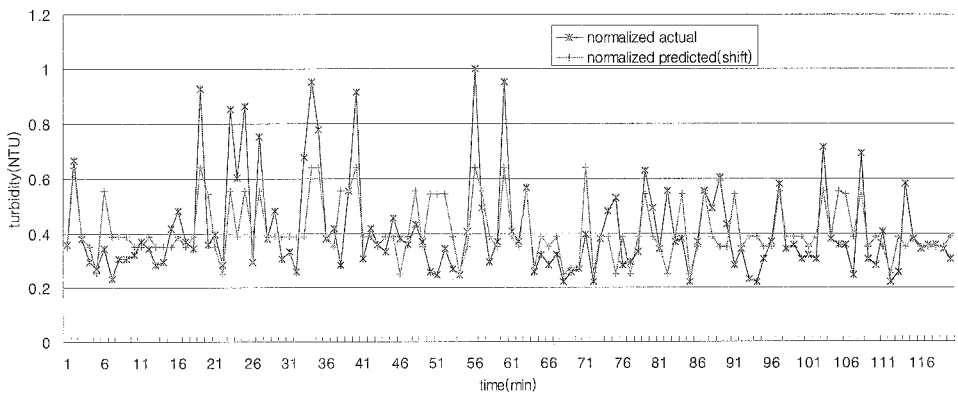


Fig. 6. 응집제 주입량, 원수의 탁도, SCD 및 conductivity를 변수로 사용하는 경우 Model tree의 검증 결과.

탁도의 경우에는 10NTU 이하의 값을 가지므로 각 LM에서 구해지는 상수 값들의 order에 영향을 미치게 된다.

결론적으로 본 연구에서 예측하고자 한 침전수의 실제 데이터는 0.18-0.81NTU의 작은 범위를 가지지만 변동 폭이 크고 급변하는 경향을 가진다. 또한 본 연구에서 사용한 운전 데이터는 대부분 약 80-490NTU의 고탁도에서 운전한 결과를 나타내고 있다. 분 단위로 변하는 원수의 수질을 기반으로 추후 공정인 침전지의 유출수 탁도를 효율적으로 예측하는 것은 기존의 해석적 모델로는 불가능한 일이다. 그러나 본 연구에서 제안하는 model tree기반의 예측 모델의 경우 외삽(extrapolation)하는 경우(학습시킨 범위를 벗어나는 경우)를 제외하고 충분한 데이터로 기계학습을 시켜 모델을 수립하면 효율적인 예측이 가능하리라 판단된다.

5. 결 론

본 연구에서는 실제 정수처리공정에서 응집 및 침전 효율에 영향을 미치는 실시간 측정 변수(시간, 응집제 주입량, 원수 탁도, SCD 및 conductivity)를 이용하여 model tree기법을 근간으로 기계학습(machine learning)시킨 floc의 침전효율을 예측하는 모델을 수립하였다. 기존의 응집효율 예측 모델은 실공정에서의 응집 및 침전 효율을 예측하고자하는 목적보다는 메커니즘을 규명하기위해 수립된 해석 모델이라 실공정상에서 발생하는 floc의 응집 및 침전 효율을 예측하기에는 무리가 있었다. 그러나 본 연구에서 수립한 model tree 기반의 응집 효율 예측 모델의 경우, 원수 수질이 급격하게 변동하는 고탁도시에도 평균 error가 9% 이하의 침전수 탁도 예측능을 보였으며, 공학적인 편의 측면에서 추가적인 인자의 측정 없이 실무에 적용할 수 있다는 장점이 있다.

상기에서 언급한 바와 같이 model tree를 이용하기 위해 변수로 고려되는 인자 중 직접적으로 응집 및 침전 효율에 영향을 미치는 인자이외에 시간 변수를 도입하는 경우, 응집 및 침전 효율의 예측능이 좋으나, 시간을 변수로서 도입하여야 할지는 아직 의문이라 할 수 있다.

참고문헌

1. 최종후 (2003). Answer Tree 3.0을 이용한 데이터마이닝 예측 및 활용, SPSS 아카데미.
2. 한국수자원공사 (2004) 반월정수장 유동전류계(SCD) 실공정 적용을 위한 연구 보고서, 한국수자원공사.
3. Dentel, S.K. and Kingery, K.M. (1988) An Evaluation of Streaming Current Detectors, Denver, CO. American Water Works Association and AWWA Research Foundation.
4. Letterman, R. D., Amirtharajah, A., and O'Melia, C.R., (1999) Coagulation and Flocculation. In *Water Quality and Treatment*, Ch.6. McGraw-Hills, New York.
5. Reynolds, T. D. and Richards, P. A. (1996) *Unit Operation and Processes in Environmental Engineering*, PWS Publishing Co., Boston, MA, pp. 166-177.
6. Walker, C. A., Kirby, J. T., and Dentel, S. K., (1996) The streaming current detector: A Quantitative model, *J. of Colloid and Interface Science*, **182**, pp.71-81.
7. Ian H. Witten, Eibe Frank, (2005) Data Mining (2nd edition), Morgan Kaufmann Publishers, New York.
8. Han, M. Y., and Lawler, D. F., (1992) The (Relative) Insignificance of G in Flocculation, *Journal of the American Water Works Association*, **84**(10), pp 79-91.
9. Smoluchoski, M., (1917) Versuch Einer Mathematischen Theorie der Koagulations-Kinetik Kolloider Losungen, *Z. Physik. Chem.*, **92**, pp 129.
10. Tambo, N., and Watanabe, Y., (1979) Physical Aspect of Flocculation Process I. Fundamnetal Treatise, *Water Research*, **12**, pp. 429-439.
11. Dimitri P. Solomatine and Yupeng Xue (2004), M5 Model Tree and Neural Network : Application to Flood Forecasting in the Upper Reach of the Huai River in China, *Journal of Hydrologic Engineering*, ASCE/November/December 2004, pp. 491-501
12. Quinlan, J. R. (1992), Learning with continuous classes. Proceeding of the 5th Joint Conference on Artificial Intelligence, Adams & Sterling, eds., World Scientific, Singapore, pp. 343-348.