데이터마이닝 로드맵 개발과 수처리 응집제 제어를 위한 데이터마이닝 적용

배현* · 김성신* · 김예진**

Development of Datamining Roadmap and Its Application to Water Treatment Plant for Coagulant Control

Hyeon Bae* · Sungshin Kim* · Yejin Kim**

이 논문은 본 연구는 산업자원부의 지원에 의하여 기초전력연구원(R-2004-B-129) 주관으로 수행된 과제임

요 약

본 논문은 정수장에서 사용하는 응집제의 종류를 결정하기 위한 시스템 개발에 관한 내용이다. 정수장은 여러 단위 처리장으로 구성되며, 불순물을 제거하기 위하여 혼화지에서 응집제를 주입하여 침전을 시킨다. 현재까지 응집제 결정을 위해 Jar-test를 이용하는데, 이 방법은 사람의 주관적인 판단에 의존하므로 실험 오차가 발생할 수 있다. 특히 정수장의 자동화를 위한 시스템 개발에서 가장 큰 걸림돌로 작용하고 있다. 본 논문은 이러한 문제점을 해결하기 위하여 로드맵에 기초한 데이터마이닝 기법을 이용하여 응집제를 선택할 수 있는 제어기를 개발하였다. 제어 규칙은 클러스터링 기법으로 도출하였는데, 군집의 초기 값과 개수는 통계적 지수 값을 사용하여 결정하였다.

ABSTRACT

In coagulant control of water treatment plants, rule extraction, one of datamining categories, was performed for coagulant control of a water treatment plant. Clustering methods were applied to extract control rules from data. These control rules can be used for fully automation of water treatment plants instead of operator's knowledge for plant control. To perform fuzzy clustering, there are some coefficients to be determined and these kinds of studies have been performed over decades such as clustering indices. In this study, statistical indices were taken to calculate the number of clusters. Simultaneously, seed points were found out based on hierarchical clustering. These statistical approaches give information about features of clusters, so it can reduce computing cost and increase accuracy of clustering. The proposed algorithm can play an important role in datamining and knowledge discovery.

키워드

Coagulant control, rule extraction, datamining, rule-based control

* 부산대학교 전자전기정보통신공학부

** 부산대학교 환경공학과

접수일자 : 2005. 8. 30

Ⅰ. 서 론

일반적인 정수처리의 공정은 응집, 침전, 여과, 살균소독 처리 과정을 거치며, 약품 주입에 의한 응집, 침전 및 살균소독 처리는 상수처리 시스템의 가장 핵심부분을 이룬다. 응집공정에서 쓰이는 응집제의 종류로는 여러 가지가 있지만, 현재 덕산 정수장에서는 PAC, PASS, PSO-M의 3가지 약품을 주로 쓰고 있다. 응집제의 선택과 주입율 결정은 원수를 Jar-test를 토대로 전문가의 경험적 지식에 의존한다. 그러나 실험을 위해 2시간 이상이 소요되므로 수시로 변하는 원수의 상황에 적절하게 대처하기 어렵다[1].

실제 정수처리에 있어서의 응집효과는 탁도, 유발물질의 양과 종류, 입자의 크기, 특정 이온의 존재여부, pH, 알칼리도 등에 영향을 받게 된다[2]. 이에 본 연구에서는 덕산 정수장의 일별수질 데이터를 참조하여 응집제 선택에 영향을 미치는 입력인자(pH, 탁도, 알칼리도, Chl-a, 수온)에 대한 규칙을 클러스터링 방법들(clustering method)을 이용하여 추출하고자 하였다.

응집제 주입에 대한 연구로 Sugeno 등은 운전자의 경험을 바탕으로 퍼지 규칙을 구성하여 제어하였다[3]. Baba 등은 운전지원시스템에 퍼지 제어를 위한 입력으로 응집결과를 영상으로 분석한 기능을 추가하였으며, Enbutsu 등은 퍼지 제어규칙을 만드는 과정에 신경망의 학습이론을 도입하여 개선된 퍼지 규칙 추출방법을 제안하였다[4]. 국내의 응집제 주입에 관한 연구를 보면 뉴로-퍼지(neuro-fuzzy)모델이나 신경망 모델을 적용한 사례들이 있다[5].

본 논문에서는 응집제 제어에 데이터마이닝 기법을 적용하여 원수수질에 대한 대처능력을 높이고, 응집공정을 자동화함으로써 전체적인 정수공정의 경제적 효율을 높이고자 하였다.

Ⅱ. 정수처리 공정에서의 응집제 제어

2.1 정수처리 공정

상수처리설비는 하천수의 취수구로부터 공급지로 급수하기까지의 정수공정을 포함한다. 정수처리 시설은 원수의 수질이 악화되더라도 음용수의 수질기준에 적합한 정수를 생산, 공급할 수 있는 기능을 갖춰야 한다.

그림 1은 정수장의 공정그림으로 취수, 오존처리, 혼화, 침전, 모래여과, 활성탄여과 순으로 처리하게 된다.

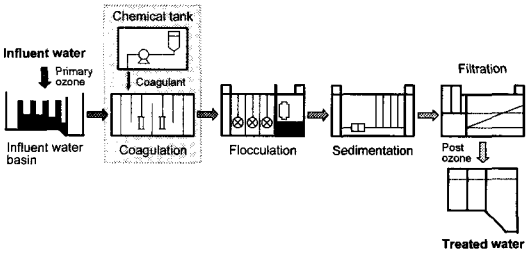


그림 1. 정수처리 공정 개요
Fig. 1 Scheme of water treatment process

2.2 응집제 종류

응집제로는 제 2세대의 응집제로 황산반토의 뒤를 이어 현재 세계적으로 앞서가는 제품 중의 하나인 PAC(Poly Aluminium Chloride), 추운겨울에 응집성능이 뛰어난 PASS(Poly Aluminium Sulfate Silicate), 그리고 갈수기 조류로 인한 pH상승 시 pH 강화효과가 있는 PSO-M(Poly Organic Aluminum Magnesium Sulfate) 등이 널리 사용된다. 각 응집제가 가지고 있는 장·단점이 있으므로 원수의 성상에 따라 선택하여 사용한다.

Ⅲ. 데이터마이닝 로드맵

본 연구에서는 데이터마이닝의 적용에서의 고민을 줄이고 쉽게 적용하기 위하여 데이터마이닝 로드맵을 구성하였다. 현재 사용되고 있는 기법들을 정리하여 목적에 따라 분류를 하여 최종 시스템 구성 방법을 보기 쉽게 도식화 하였다. 기존의 연구에서는 카테고리를 나누어 분류를 하였지만 그 다양성이 제한되었고 주로 마케팅과 같은 분야에서의 분류를 고려하여 생산에서 필요한 데이터를 다루는 로드맵은 찾기가 쉽지 않다. 본 논문에서는 이러한 생산에서 필요한 데이터마이닝 적용을 위한 로드맵을 제안하였다. 본 논문에서 개발된 데이터마이닝 로드맵은 그림 2에서 보는 바와 같이 크게 전처리부, 검출부, 분류 및 진단부로 구성된다. 각 단계에서 적절한 기능을 수행함으로써 전반적으로 안정된 시스템을 구성할 수 있다.

3.1 데이터 전처리 단계

전처리부는 그림 2에서 보는 바와 같이 데이터마이닝 기법을 적용하기에 앞서 알고리즘의 성능을 높이기 위해 필요한 데이터 전처리를 수행하는 단계이다. 제조 현장에서 측정되어 수집된 데이터는 원하는 마이닝 결과를 유도하기에 부적절한 경우가 빈번하다. 따라서 데이터마이닝의 목적에 따라 적절한 전처리가 수행되어야 한다. 전처리부는 크게 데이터 정제, 데이터 변환, 그리고 데이터 생성으로 나뉜다. 본 논문에서는 데이터 정제 과정이 수행되었다.

3.2 데이터마이닝 및 지식 획득 단계

데이터마이닝과 지식 획득에서는 그림 2와 같이 가공된 데이터로부터 원하는 특징을 찾는 과정을 수행한다. 데이터마이닝에서 가장 중요한 부분으로 다양한 기법들이 사용되고 있다. 기존의 연구들에서는 주로 마케팅과 같은 오프라인에 대한 데이터마이닝 범위를 분류하여 적용에 있어서의 다양성이 떨어졌다. 또한 다양한 내용 중 어떤 카테고리를 적용할 것인지를 선택하는 것은 쉽지 않은 과정이다. 특히 사례에 대한 경험이 많지 않은 경우 숱한 시행착오를 거쳐야 하는 어려움이 있다. 마케팅 데이터와 같이 오프라인 데이터는 잘못된 적용의 경우 그 파급이 크지 않을 수 있지만 생산 현장과 같은 온라인 라인에의 적용 시 잘못된 데이터마이닝의 적용은 큰 문제를 유발할 수 있다.

본 연구에서는 이러한 문제점을 해결하기 위하여 각 카테고리를 상세히 나누었다. 각 카테고리는 몇 가지의 검증된 기법을 목적에 따라 분류되어 있으므로 목적에 따라 카테고리를 결정하게 되면 기법들은 쉽게 결정이 된다. 본 단계는 특징 추출, 분류, 예측, 규칙 도출, 그리고 클러스터링 카테고리를 포함한다. 본 논문에서는 약품 종류 선택을 위해 퍼지 c-means를 이용하여 클러스터링을 수행하여 규칙을 도출하였다.

3.3 시스템 구성 단계

본 논문에서는 그림 2에서 보는 바와 같이 데이터마이닝의 최종 단계를 두 개의 큰 시스템 구성에 있다고 제안하였다. 먼저 규칙을 통한 추론 시스템이 있고 나머지 하나는 모델을 통한 예측 시스템이 있다. 두 시스템을 통합하여 필요에 따라 하나의 전체 시스템을 구성하게 된다.

기존의 데이터마이닝 관련 문헌에서는 단위 기법에 대한 분류에 대한 언급은 있었지만 최종적인 시스템으로의 구성에 대한 언급은 없었다. 본 연구에서 제안한 로드맵에서는 데이터마이닝 및 지식 획득을 통해서 구성되는 결과를 크게 두개의 시스템으로 나누었다. 추론 규칙과 데이터 모델이 그것이다. 본 연구에서는 추론 규칙의 도출을 적용하였다.

3.4 제안된 데이터마이닝 로드맵

그림 2는 본 논문에서 제안한 데이터마이닝 로드맵을 보여주고 있다. 본 연구에서 좌측에 보이는 단계들이 큰 범주를 보여주고 있고, 우측의 그림에서 각 세부적인 범주를 보여 주고 있다. 큰 범주 안에 사용되고 있는 방법들을 함께 표시하여 사용자가 접근하기 쉽도록 하였다. 각 방법들은 현재까지 널리 사용되는 방법을 적절하게 선택하여 구성하였다. 보다 많은 방법이 포함될 수도 있을 것이다.

데이터마이닝 로드맵의 목적은 사용자가 데이터로부터 목적을 달성하기 위해 필요한 과정과 방법을 선택하는데 도움을 주기 위해서이다. 많은 경험을 가지지 않은 사용자도 목적만 뚜렷하다면 결과까지 쉽게 이끌 수 있을 것이다. 본 로드맵의 가장 큰 목적이 데이터로부터 원하는 정보를 추출하는 과정을 적절하게 수행하도록 도움을 주는 것이다. 반면 포함된 각 방법들은 적용에서 세부적인 조정이 필요하다. 이러한 조정은 사용자가 직접 적용하면서 해결해야 하는 부분이므로 로드맵에서 구체적으로 언급하기는 어렵다. 그렇지만 큰 과정과 방법이 정의됨으로서 시행착오의 과정을 줄일 수 있는 장점을 가지고 있다.

Ⅳ. 실험 결과 및 검토

4.1 원수데이터 특성

본 연구에서 사용한 데이터는 덕산 정수장에서 취득한 것으로 응집제 선택은 수질데이터의 pH, 탁도(Tu), 알칼리도(Al)에 따라 결정되며 pH와 알칼리도(Al)는 참조자료로 쓰인다. 탁도가 투입량에 큰 비중을 차지하는데 50NTU이상에서는 PAC를 사용하나 이하에서는 조류가 심할 때는 PASS를 사용하고, 그렇지 않을 때는 PSO-M을 사용한다. 현장에서 응집제의 선택

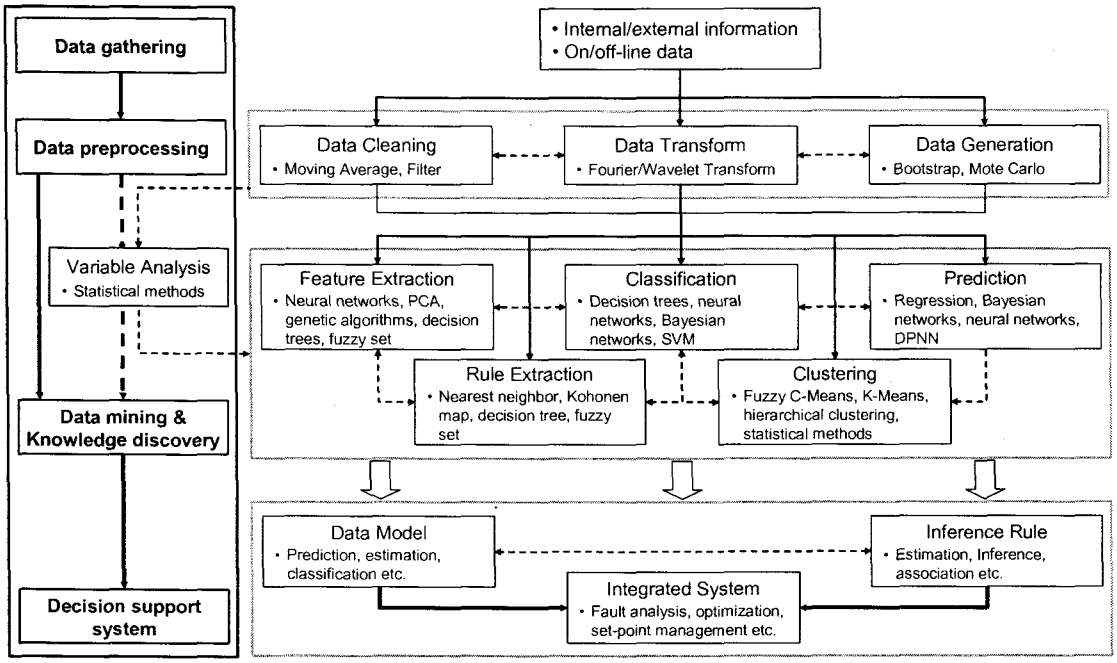


그림 2. 제안된 데이터마이닝 로드맵
Fig. 2 Proposed datamining roadmap

에 중요한 요소로 작용하는 pH, 탁도, 알칼리도, 클로로필, 수온을 클러스터링의 입력으로 사용하였다.

이와 같은 원수 조건으로부터 적절한 약품의 종류를 선택하기 위하여 데이터마이닝을 적용하였다. 적용된 방법들은 앞서 언급한 데이터마이닝 로드맵으로부터 선택되는데, 본 연구의 목적이 약품의 종류를 선택하기 위한 로직(logic)이므로 규칙 도출 카테고리가 사용된다. 이렇게 도출된 규칙은 응집제 종류를 결정하는 규칙으로 표현된다.

4.2 클러스터링을 이용한 규칙 도출

규칙 도출을 위하여 다양한 방법들이 적용되고 있다. 본 논문에서는 비선형 특성과 경계 값에서의 분류 장점을 위해 클러스터링 방법을 사용하였다. 기존의 방법에서 문제시 되었던 초기 값 결정과 개수 결정을 위한 방법을 통계적인 기법을 이용하여 해결하였다.

초기 클러스터의 시드 값은 계층적 군집분석을 이용하여 구한다. 군집의 개수가 정해지고 나면 그 개수에 따른 군집을 나열한 후 각 클러스터에 포함된 개체의 중심을 구함으로서 초기 시작 값을 결정할 수 있다. 본 연구에서는 각 계층적 방법을 적용하여 초기 값을 구한 후 값의 평가를 통해 적절한 값을 선택하여 사용한다.

클러스터의 개수를 구하는 것은 퍼지 c-means에서 중요시되고 있다. 앞서 언급하였듯이 많은 연구자들이 각각의 지수들을 제안하였다. 본 연구에서는 통계에서 사용하는 계층적 클러스터링 방법을 이용하여 적절한 클러스터 개수를 결정한다. 각 지수들은 일반적인 통계 값 계산 지수들이다. 그림 3과 4는 여러 지수 중 근평균 자승 편차와 R-자승의 결과를 보여주는데 값의 변화가 큰 클러스터 개수가 적절한 개수를 의미한다. 본 연구에서는 3~5개의 클러스터가 후보군으로 결정되었다. 이 중 최종 클러스터 개수는 규칙의 분류 성능에 의해 결정되는데, 본 연구에서는 4개의 클러스터로 나누어 규칙을 생성한 경우 성능이 가장 우수하였다. 그림 5는 탁도와 chl-a 변수에 대하여 4개의 군집 개수로 군집화한 결과를 보여주고 있다. 군집 결과를 규칙으로 정리한 것이 표 1과 같다.

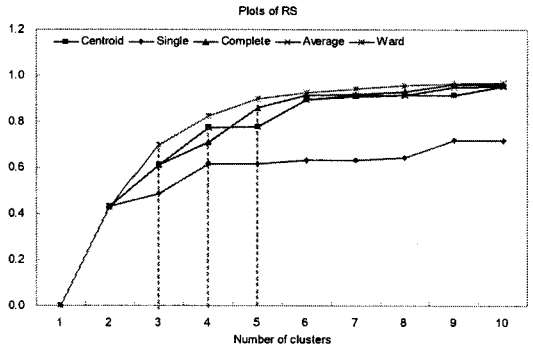


그림 3. RS 지수 결과를 이용한 개수 선택
Fig. 3 Selection of cluster numbers using RS

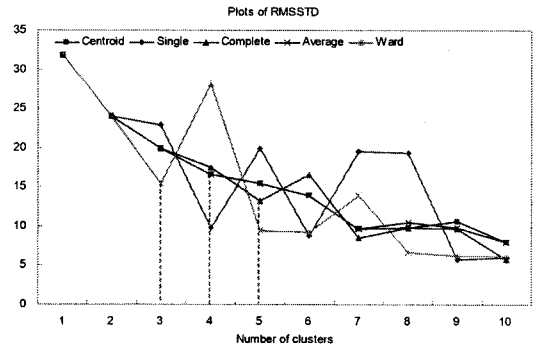
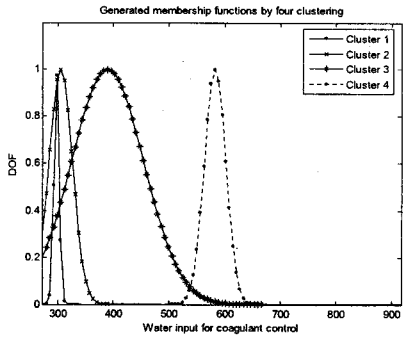
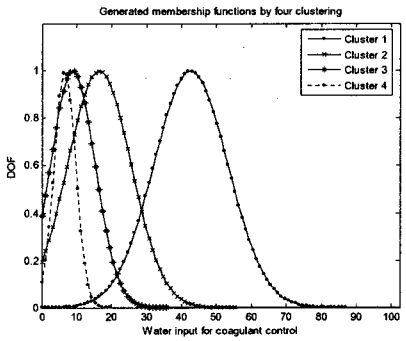


그림 4. RMSSTD 지수를 이용한 개수 선택
Fig. 4 Selection of cluster number using RMSSTD



(a) 탁도에 대한 클러스터링 결과



(b) Chl-a에 대한 클러스터링 결과
그림 5. 퍼지 클러스터링에 의한 규칙 도출
Fig. 5 Result of rule extraction using fuzzy clustering

· 표 1. 클러스터링 결과로부터 도출한 규칙
Table. 1 Extracted rules from clustering results

If Temperature is L, Turbidity is L, Alkalinity is H, and **Chl-a is H**, Then Coagulant is PASS.
If Temperature is H, **Turbidity is H**, Alkalinity is L, and Chl-a is L, Then Coagulant is PAC.
If Temperature is M, Turbidity is M, Alkalinity is M, and Chl-a is M, Then Coagulant is PSO-M.

V. 결 론

적절한 응집제와 주입율의 결정은 처리효율과 수질의 개선을 가져온다. 따라서 응집반응에 영향을 미치는 주요 수질인자의 변화에 신속하고 정확하게 응집제의 선택과 주입율 결정을 위한 방법으로 본 연구에서는 응집제 선택 시에는 클러스터링 방법을 이용하였으며, 응집제 주입율은 저자의 사전 연구에서 신경회로망으로 검토된 결과가 있으므로 생략하였다. 본 연구에서 얻어진 응집제 선택 규칙을 적용·평가하기위하여 실제 정수처리 시설에서의 검증이 필요하다고 판단된다.

감사의 글

이 논문은 본 연구는 산업자원부의 지원에 의하여 기초전력연구원(R-2004-B-129) 주관으로 수행된 과제임.

참고문헌

[1] 수자원연구소, 응집제 투입량 자동결정 시스템 개발연구, 한국수자원공사 기술보고서, pp. 14-17, 1977.

[2] A. P. Black and S. A. Hamnah, "Electrophoretic Studies of Turbidity removal Coagulant with Aluminum Sulfate," J. AWWA, vol. 53, pp. 438, 1961.

[3] O. Yagishita, O. Itoh, and M. Sugeno, "Application of fuzzy reasoning to the water purification process," Industrial Application of Fuzzy Control, pp. 19-39, 1985.

[4] K. I. Enbutsu, K. Baba, N. Hara, K. Waseda, and S. Nogita, "Integrationof multi AI paradigms for intelligent operation support systems-fuzzy rule extraction from a neural network," IAWQ, pp. 333-340, 1993.

[5] 배현, 김성신, 최대원, 이승태, 김예진, "원수조건에 따른 상수도 응집제 종류와 주입량 결정을 위한 데이터 마이닝 적용," 한국퍼지 및 지능시스템학회 논문지, vol. 15, no. 1, pp. 1-5, 2005. 2. 25.

저자소개

배현(Hyeon Bae)

1999. 2. 경상대 전기공학과 공학사
2001. 2. 부산대 전기공학과 공학석사
2005. 8. 부산대 전기공학과 공학박사
2005. 9.~현재 부산대 전기공학과 박사후연구원

※관심분야 : 데이터마이닝, 지능제어 및 시스템, 생명정보학 및 시스템 생물학

김성신(Sungshin Kim)

1984. 2. 연세대학교 전기공학과 공학사
1986. 2. 연세대학교 전기공학과 공학석사
1996. 8. Georgia Institute of Technology 공학박사
1998. 3.~현재 부산대 전자전기정보통신공학부 부교수

※관심분야 : 지능제어 및 시스템, 데이터마이닝, 생명공학, 시스템 생물학

김예진(Yejin Kim)

2000. 2. 부산대 환경공학과 공학사
2002. 2. 부산대 환경공학과 공학석사
2003. 8.~현재 부산대 환경공학과 박사과정

※관심분야 : 수처리공학, 환경모델링, 공정 제어 및 진단