물 수요예측을 위한 데이터 마이닝 기법 분석

신강욱, 홍성택
한국수자원공사 K-water연구원

Data mining analysis for short-term water demand forecasting

Gang-Wook Shin, Sung-Taek Hong
Korea Water Resources Corporation, KIWE

Abstract - 본 연구에서는 안정적인 물 공급과 에너지의 효율적 사용을 위한 단기 물 수요예측에 대하여 데이터 마이닝 기법의 적용성을 검토하고자 한다. 물 공급이 이루어진 요일과 특이일에 대한 시계열 분석을 통한 단기 물 수요예측과 데이터 마이닝 기법을 적용한 결과를 상호 비교하여 데이터 마이닝 기법의 적용성을 제시하고자 한다. 이를 통하여 단기 물 수요예측알고리즘의 실용화 가능성을 높일 뿐만 아니라 실시간 예측을 위한 기초 데이터 마이닝 체계를 구축하고자 한다.

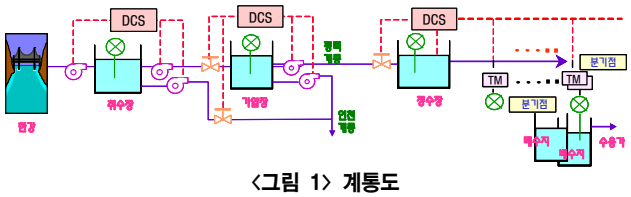
1. 서 론

수요예측 분야는 경제성장, 시스템 보안, 그리고 경영 및 계획 등 많은 응용분야에 있어서 근간이 되어왔다. 많은 응용분야 중 물 수요예측에 있어서는 두 가지 중요 인자에 의하여 특히 의존성이 높다할 수 있다. 하나는 낮과 밤 동안의 인간 사회활동과 습관에 의한 것, 그리고 나머지 하나는 날씨 조건에 따른 영향을 나타낸다. 이러한 두 가지 중요 인자에 대한 조사를 통하여 다양한 물 수요예측알고리즘을 개발하려는 시도가 있어왔다. 물수요예측을 기간별로 크게 구분하면, 장기예측과 단기예측으로 나눌 수 있다. 장래의 수도시설의 건설 혹은 확장계획 등의 년간계획을 결정하는 일을 목적으로한 물 수요예측을 장기예측으로 한다. 한편, 수도시설의 합리적인 물 운용과 유지관리계획을 목적으로 하루 또는 시간계획의 물 수요예측을 단기 물수요예측이라 한다. 특히 국토개발이나 도시계획에 의한 인구 증감에 따른 장기 물수요예측은 일반적으로 양호한 예측결과를 도출하였다. 그러나 안정적인 물 공급과 에너지의 효율적 사용을 위한 일별 및 시간별 단기 물수요예측에 대한 연구는 간헐적으로 진행되었지만, 실제 적용할 수 있는 수준의 결과도출이 미흡한 실정이다. 지금까지 단기 물수요예측을 위하여 ARIMA 모델을 비롯한 뉴럴알고리즘, 그리고 칼만필터등 다양한 시도가 있었다[1-3]. 본 연구에서는 단기 물 수요예측을 위한 시계열분석을 통하여 데이터 마이닝 기법의 적용성을 고려한 알고리즘을 제시하고자 한다.

2. 본 론

2.1 개요

본 연구에서 검토된 대상 사업장은 일일 공급 가능한 급수이 약 100만톤이며 급수 인구는 약 200 만명인 정수장이다. 대상 정수장은 그림 1에서와 같이 취수장으로부터 유입된 유량을 정수장에서 정수처리 후 배수지 혹은 수용가로 공급하도록 구성되어 있다. 각 배수지 및 수용가에서의 물 사용량의 합에 의한 예측의 어려움으로 인하여 정수장 유출 단계에서의 통합 유량을 통한 단기 물 수요예측알고리즘을 도출하고자 한다.



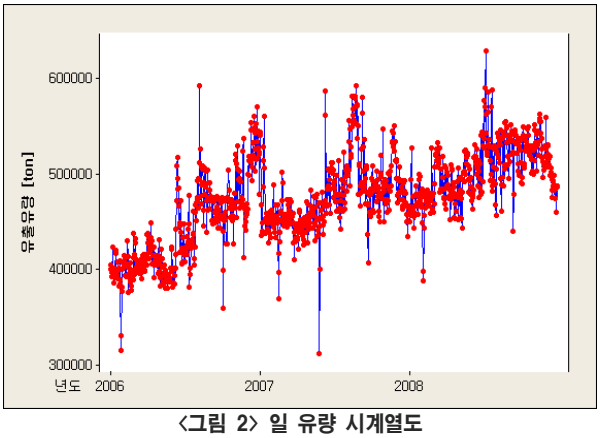
<그림 1> 계통도

2.2 물 공급량 분석

A 정수장에서의 년간 물공급량을 분석하기 위하여 2006년부터 2008년까지 3년 동안의 물 공급데이터를 조사하였다. 그림 2는 일일 공급량에 대한 시계열도를 나타낸 것이며, 시계열도에서 알 수 있는 바와 같이 여름철에서의 물 공급량이 대체로 많은 것을 알 수 있다. 또한, 년도별 평균 공급유량을 살펴보면 2006년도 44만톤, 2007년도 47만7천톤, 그리고 2008년도 약 50만톤의 물이 공급되었다. 이와 같이 매년 평균적으로 약 3만톤 규모의 공급량이 증가된 것을 알 수 있다. 이를 추세선으로 나타내면 다음 (1)식과 같이 1차 선형식으로 표현할 수 있다.

$$Y = 418333 + 100.66 * t \quad (1)$$

여기서, Y는 용수 공급량, t는 일단위이다.
따라서, 식 (1)에서 알 수 있는바와 같이 하루에 약 100 톤 규모의 물 공급량 증가가 있음을 가정할 수 있다.



<그림 2> 일 유량 시계열도

2.2.1 요일별 유량 특성분석

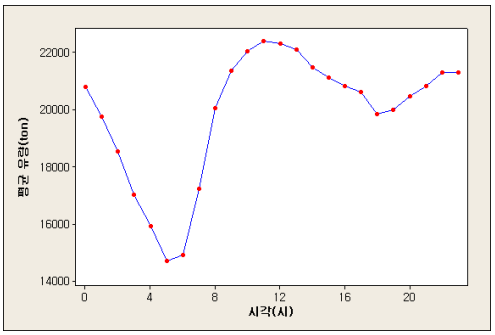
3년간의 물공급 데이터에 대하여 요일별 일원분산분석과 특이일에 대한 일원분산분석을 실시하였다. 표 1에서와 같이 평일과 주말, 그리고 특이일에서의 평균값이 비교적 많은 차이가 있음을 알 수 있다. 또한, 특이일에서의 휴일 전후에 대한 물 사용량에 있어서는 평일의 물 사용량과 별 차이가 없음을 알 수 있다.

<표 1> 요일별 일원분산 분석

수준	N	평균	표준 편차
월	153	479293	43765
화수목	458	476389	47967
금	151	476699	46080
토	151	470235	42019
일	152	467451	40845
휴일전	6	474982	43351
설추석	19	416817	53819
휴일후	6	446332	48498

2.2.2 시간대별 유량 특성분석

계절적인 변화에 대하여 고려하지 않고 3년간의 시간대별 공급유량에 대하여 분석하였으며. 시간대별 그래프는 그림 3과 같다. 그림에서 알 수 있는바와 같이 새벽 0시부터 5시까지 물 공급량이 감소하는 것을 알 수 있다. 또한 낮 12시경 물 소비량이 가장 많은 것으로 나타나 일일 물 공급량의 최대치를 나타내었다.



〈그림 3〉 시간대별 일원분산 분석

2.3 다중자기회귀 분석

2.3.1 상관계수 선정

다중자기회귀 분석을 위하여 2006년과 2007년도 기상데이타와 공급유량 데이터를 이용하여 상관관계를 분석하였다. 수집된 기상데이타는 최고기온, 평균기온, 최저기온, 평균풍속, 상태습도, 일조시간, 운량, 강수량, 적설량 등이다. 공급유량과의 상관계수를 구한 결과 표 2와 같은 결과를 얻었으며 본 연구에서는 상관관계가 높은 평균기온을 매개변수로 선정하였다.

〈표 2〉 기상데이타 상관계수

기상 변수	최고기온	평균기온	최저기온	평균풍속	운량
상관계수 값	0.304	0.316	0.320	-0.087	0.019
기상 변수	상태습도	일조시간	강수량	적설량	전일유량
상관계수 값	0.124	-0.021	0.013	-0.009	0.742

2.3.2 다중자기회귀 모델

일 공급유량을 예측하기 위하여 주요 상관계수를 평균기온과 전일유량으로 선정하였으며, 요일 특성별 다중자기회귀 모델을 도출하여 각각 분석을 실시하였다. 2006년도와 2007년도 데이터를 이용하여 다중자기회귀 모델식을 도출하였으며 이의 추정오차를 산출하였다. 도출된 다중자기회귀식을 이용하여 2008년도 데이터에 적용하여 예측오차 결과를 산출하였으며 표 3과 같다. 표 3에서 알수 있는바와 같이 요일별 일원분석 결과와 요일별 중회귀 분석 결과에서는 적용성이 상이함을 알 수 있었다. 여기서 추정오차와 예측오차는 각각 평균절대 백분위 오차(MAPE:Mean Absolute Percentage Prediction Error)를 나타낸다.

일 공급유량에 대한 다중자기회귀 모델식은 다음과 같다.

〈표 3〉 다중자기회귀 모델식 오차특성

구분	중회귀식	추정오차(%)	예측오차(%)
월	126421+454*T+0.72*Q	4.58	4.0
화수목금	135710+362*T+0.659*Q	3.02	6.65
토	118582+415*T+0.73*Q	4.25	3.8
일	176493+200*T+0.608*Q	4.59	4.38
휴일	3676271+2697*T-0.016*Q	8.59	8.96
월화수목금	78295+312*T+0.823*Q	3.78	3.48
토일	110875+148*T+0.754*Q	4.04	3.51
6,7,8월	37527+3838*T+0.726*Q	3.82	4.56
여름외 기간	69002+81*T+0.846*Q	3.62	3.08
모든일	70072+234*T+0.841*Q	3.73	3.35

2.4 데이터 마이닝 분석

2.4.1 데이터 마이닝 기법

데이터 마이닝(data mining)은 대규모 데이터 저장소에서 유용한 정보를 자동적으로 탐색하는 과정이다. 데이터 마이닝 기법은 대규모 데이터베이스를 구석구석 뒤져서 모른 채 넘어갈 수 있는 새롭고 유용한 패턴을 탐색하기 위해 적용된다[4].

데이터 마이닝의 핵심 작업은 다음 4가지로 나타낼 수 있다.

첫째, 예측 모델링(predictive modeling)은 목표 변수를 설명 변수의 함수 모델로 생성하는 작업으로서, 이산형 목표 변수에 사용하는 분류(classification)와 연속형 목표 변수에 사용하는 회귀(regression)의 두 가지 유형으로 나눌 수 있다.

둘째, 연관 분석(association analysis)은 데이터에 강하게 연관된 특징을 설명하는 패턴을 발견하는 데 사용한다.

셋째, 군집 분석(cluster analysis)은 동일한 군집에 속하는 관측들은 다른 군집에 속하는 관측보다 더 유사하도록 긴밀하게 관련된 관측의 그룹을 탐색하는 것이다.

넷째, 이상치 탐지(anomaly detection)는 특징이 다른 나머지 데이터들과 현저히 다른 관측들을 식별하는 작업이다.

2.4.2 이상치 탐지

데이터 마이닝 기법의 도입을 통한 다중자기회귀 모델식 도출을 위하여 이상치 탐지를 실시하였다. 이상치 탐지의 기준은 전일 유량대비 10 % 범위를 벗어난 경우 이상치로 설정하였으며 이를 통하여 다중자기회귀 모델식을 다음 (2)식과 같이 도출할 수 있었다.

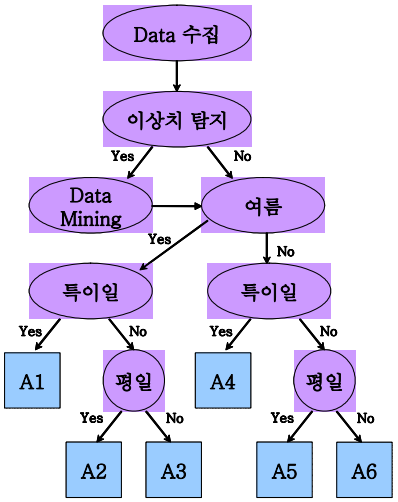
이상치 탐지한 후 10 % 범위를 벗어난 경우 전전일의 유량값을 대입하였으며 이에 따른 추정오차는 기존의 3.73 %에서 2.92 %로 상당히 낮아졌다. 또한 예측오차를 도출하기 위하여 기존의 2008년 원시데이타에 적용한 결과 3.35 %에서 3.29 %로 다소 낮아진 결과를 얻었다. 2008년의 원시데이타를 이상치 탐지를 통하여 데이터 마이닝한 후의 데이터를 이용한 예측오차는 2.94 %로 나타났다.

$$Qp = 41992 + 146*T + 0.905*Q \qquad (2)$$

여기서, Qp는 예측 유량이며, T는 평균기온, Q는 전일 유량을 각각 나타낸다.

2.4.3 의사결정 트리기법 적용

계절별, 월별, 요일별, 그리고 시간대별 특성분석을 통하여 단기 물수요예측을 위한 의사결정 트리기법을 적용하기 위한 방안은 그림 4와 같다. 계절 특성 중 여름특성, 그리고 특이일 및 평일과 주말에 대한 부분으로 크게 분류하였다. 이는 세분화된 트리의 경우 유사 특징으로 인하여 오히려 시스템의 안정성을 저해할 수 있기 때문이다.



〈그림 4〉 의사결정 트리기법 적용을 위한 개념도

3. 결 론

본 연구를 통하여 요일별 특성을 포함한 계절적 특성 등에 대한 상관계수를 도출하였으며, 이를 이용하여 다중자기회귀 분석을 실시하였다. 특히, 데이터 마이닝 기법을 적용하여 일일 유량의 평균절대 백분위 오차가 추정단계에서는 3.73 %에서 2.92 %로, 예측단계에서는 3.35 %에서 3.29 %로 양호해 졌음을 확인하였다. 향후에는 의사결정 트리기법의 적용성을 높일 수 있는 방안에 대하여 연구하고자 한다.

[참 고 문 헌]

[1] H.M. Al-Hama, S.A. Soliman, “Short-term electric load forecasting based on Kalman filtering algorithm with moving window weather and load model”, Electric power systems research, No. 68, pp. 47-59, 2004.

[2] 한태환, 남의석 “칼만필터의 적응형모델 기법을 이용한 광역상수도 시스템의 수요예측 모델 개발”, 한국조명전기설비학회, Vol. 15, N0. 2, pp. 38-48, 2001.

[3] 김신걸,변신숙,김영상,구자용,“시스템 다이내믹스법을 이용한 서울특별시의 장기 물수요예측”, Vol. 20, No. 2, pp. 187-198, 2006

[4] Pang-Ning Tan, "Introduction to Data Mining”, 2006