NIR 분광자료를 이용한 원유 수율에 대한 예측모델 개발

박광수* · 고영현* · 이혜선* · 전치혁*

* 포항공과대학교 산업공학과

Abstract

최근 근적외선(near infrared; NIR) 분광분석을 이용한 물성의 정량적인 분석을 실시할 때에 다양한 다변량 분석방법이 시도되고 있으며 그 응용 역시 다양한 분야에 걸쳐서 적용되고 있다. NIR 자료를 이용한 분석에서는 입력 변수간 다중 공선성의 문제 및 관측치 보다 더 많은 입력변수의 문제가 있다. 이러한 문제를 해결하기 위해서는 주로 두 가지의 방법을 택하고 있다. 분광분석에 의한 결과는 하나의 관측치에 대해 파장별로 연속된 자료 형태를 가지기 때문에 인접한 파장과는 아주 큰 상관관계를 가진다. 따라서 이들 상관관계를 피할 수 있는 방법으로 기존의 상관관계가 큰 변수들 중 상관관계가 적은 변수조합을 선택하는 것이 첫번째 해결 방법이고, 두번째는 선형변환(linear transform)과정을 거쳐서 나오는 주성분(principal component)을 이용하는 것이다.

변수선택 방법으로는 회귀 분석의 stepwise방법을 사용하기도 하며 최적 변수 조합을 선택하기 위하여 신경망, 유전자 알고리듬 등의 다양한 방법이 이용되기도 한다. 주성분을 이용하는 대표적인 분석방법으로는 부분최소자승회귀분석 (partial least square regression; PLSR)과 주성분 회귀 분석 (principal component regression; PCR)이 있다. NIR자료의 분석시 전통적인 회귀 분석 (multiple linear regression; MLR)을 이용하면 입력 변수들간의 다공선성 (colinearity)의 문제가 발생한다. 분석 자료의 관측치 수는 제한 되어 있으면서 상대적으로 입력 변수의 수가 크기 때문에 변수축소 문제가 주요한 해결 방법이다. PCR과 PLSR등의 다변량 분석기법은 이러한 두 문제를 해결하는 방법으로 제안된 대표적인 선형 모형이다. 새로운 입력값이 되는 주 성분 변수는 회귀 모형 수립시 변수간 독립을 보장 할 수 있으며 또한 사용되는 주성분의 수만큼의 적은 수의 변수로도 기존의 입력 변수들이 가지는 변화를 대부분 수용할 수 있어서 변수 축소의 효과를 가지면서도 반응변수인 목적값의 설명정도를 유지할 수 있다.

NIR 분광 자료를 이용한 다변량 분석시 이용되는 PCR과 PLSR모델은 선형 모델이라는 한계를 가지고 있으며, 자료에 감추어져 있는 비선형성을 무시하게 되는 결과를 낳는다. 따라서 본 연구에서는 이러한 비선형성을 포함할 수 있는 대표적인 분석기법인 인공 신경망(artificial neural network; ANN)

을 이용하여 원유(crude oil)의 수율 예측 모델 개발에 응용하였다. 다양한 다변량 분석을 위한 연구에서 신경망은 변수선택 및 물성예측 등의 모델 수립에서 응용되고 있지만 본 연구에서는 예측모델 수립에 응용하였다. 신경망은 그 특성상 활성함수 (activation function)의 선택에 따라 기존의 회귀 분석과 같은 결과를 내는 선형 모델의 수립에도 이용되지만 비선형 활성함수의 선택에 의해서 자료속에 포함되어 있을 수 있는 다양한 비선형성까지 분석할 수 있는 장점이 있다.

본 연구에서는 주어진 NIR 분광자료를 이용하여 다음의 두가지 변수 축소 방법을 거친 후, 이들을 신경망의 입력자료로 하는 신경망 예측모형을 수립하였다. 첫째는 회귀분석에서 stepwise방법을 통하여 적당한 수의 변수 조합으로 입력변수 수를 축소한 다음 이 자료를 신경망의 입력값으로 사용하여 모델을 수립하였으며, 두번째는 주성분 분석 과정에 결정되는 일정 수만큼의 주성분을 이용한 score값을 신경망의 입력변수로 사용하였으며 또한 PLSR의 분석시에 중간 결과값으로 나오는 PLS score를 신경망의 입력값으로 사용하였다.

회귀분석 방법을 이용한 변수 축소방법은 NIR 자료가 가지고 있는 다중공선성 및 과다변수수의 문제를 어느 정도는 해결하여 주는 방법이며 또한 반응변수의 목표값이 되는 물성을 잘 설명하는 파장을 직접적으로 파악 할 수 있다는 장점이 있지만, 입력변수들간의 상관관계를 완전히 제거할 수 없는 단점을 가지기 때문에 예측모델 수립에 있어서 여전히 추정 모수들의 신뢰도를 떨어뜨리는 결과를 낳는다. 반면에 PCR 혹은 PLSR의 score를 이용한 분석에서는 입력 자료가 가지는 변화정도 및 예측력의 대부분을 반영하면서도 일정수의 축소된 변수만을 이용하면서도 변수들간의 완전 독립을 보장할 수 있다. 하지만 기존의 자료를 선형 변형 시킨 결과를 다시 이용하기 때문에 목표치인 물성을 설명하는 파장을 쉽게 파악할 수 없다는 단점을 가진다.

본 연구에서는 원유의 NIR분광 결과로 구해지는 69개의 관측치가 분석자료로 사용되었으며 각관측치는425개의 파장 변수를 가진다. 원유의 예측분석 물성으로는 9단계 온도별 원유의 증류 수율과 API 값이 사용되었으며 이들을 예측하는 PCR, PLSR 모델 및 신경망 모델을 수립하여 그 결과를 비교하였다.

결과적으로는 완전한 다중공선성을 제거하지 못한 stepwise방법보다는 주성분 분석에 의한 score를 이용한 신경망 모델이 더 좋은 결과를 보여준다. 또한 주어진 자료를 모두 이용하는 분석에서 기존의 선형 모델인 PLSR혹은 PCR보다는 본 연구결과 신경망 모델이 다소 더 좋은 결과를 내고 있다. 특히 PLS의 score를 이용한 신경망 모델은 기존의 PLSR이 가지는 특징인 입력 변수들간의 상관관계 뿐만 아니라 입력변수와 반응변수들과의 관계를 이미 고려한 것이기에 그 결과가 PCR의 score를 이용한 것보다 더 좋게 나왔다. 이것은 다른 대부분의 다변량 분석에서 PCR과 PLSR의 비교 분석 결과와 일치하는 것으로 score를 이용한 신경망 모델에서도 유사한 결과가 구해진다. 또한 신경망 모델이 선형 예측모델 중 가장 좋은 결과를 보여주는 PLSR보다는 다소 좋은 결과를 보여준다. 이는 신경망이 이미 선형성을 포함하면서도 비선형성까지 분석을 실시하고 있기 때문이다.