김일성종합대학학보
(자연과학)
주체106(2017)년 제63권 제8호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 63 No. 8 JUCHE106(2017).

다변량검량모형작성에서 유전알고리듬에 의한 최적검량시료와 최적분석파장의 동시선택

최강진, 박영길, 리주칠

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《기초과학을 발전시키는데도 힘을 넣어야 합니다. 기초과학을 발전시키지 않고서는 인민경제 여러 부문에서 나서는 과학기술적문제를 원만히 풀어나갈수 없습니다.》(《**김정일**선집》 증보판 제11권 138페지)

다변량검량모형작성에서 검량시료와 분석파장에 대한 스펙트르초기자료행렬로부터 목적정보를 그대로 포함하면서 배경정보가 삭제된 자료행렬을 얻으면 검량모형작성에 드는 품을 줄일뿐아니라 검량모형의 성능을 개선할수 있다. 그러나 현재까지 초기자료행렬에서 최적시료모임이나 최적분석파장을 개별적으로 선택하는 방법[2－4]에 대하여서는 많이 알려져있으나 분석시료와 분석파장을 동시에 선택하는 방법은 적게 연구되였다.

우리는 다변량검량모형작성에서 최량탐색방법인 유전알고리듬(GA)[1]을 리용하여 최적검량시료와 최적분석파장을 동시에 선택하는 방법을 새롭게 제기하였다.

1. 유전알고리듬에 의한 최적검량시료와 최적분석파장선택원리

검량시료와 파장을 동시에 선택하여 전체 검량시료에 대한 스펙트르정보와 화학정보를 최대로 반영하는 자료행렬을 구성하기 위한 유전알고리듬의 원리도는 그림 1과 같다.

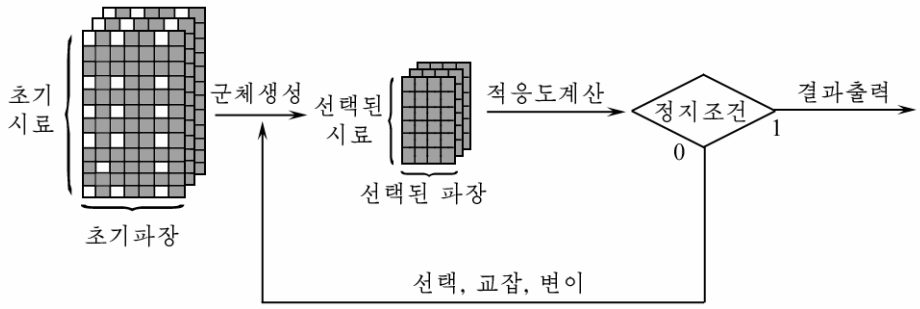

그림 1. 최적검량시료와 최적분석파장선택을 위한 유전알고리듬원리도

부호화단계에서는 2진부호화방법을 리용하여 해당 검량시료 혹은 분석파장이 선택되는 경우 1, 선택되지 않은 경우 0을 취하도록 부호렬을 구성한 다음 검량시료부호렬과 분석파장부호렬을 하나의 부호렬로 련결시킨다.

또한 교잡연산자를 부분적으로 수정하여 부호렬에서 검량시료부분과 분석파장부분이 서로 교잡되는 현상이 나타나지 않도록 하였다.

선택, 변이 등 다른 연산자들은 표준유전알고리듬[1]의 연산자를 그대로 리용하였으며 파라메터들은 표 1과 같이 설정하였다.

표 1. 유전알고리듬의 파라메터설정

구 분	값
매 염색체에 할당되는 변수개수	2(선택/취소의 2진부호화 리용)
회귀방법	다중선형회귀, 주성분회귀, 부분최소두제곱법회귀
응답자료	교차검정에서의 평균2차뿌리예측오차(RMSEP)
교잡확률/%	80
변이확률/%	10
실행회수	300

2. 시료 및 실험자료준비

1) 시료준비

켈달법으로 단백질함량을 측정한 136개의 밀가루시료들에 대하여 근적외선분광기(《S400》)로 1 300～2 100nm구간에서 16nm의 분해능으로 3회 반복하여 스펙트르를 측정하였다. 이때 얻은 근적외선흡수스펙트르들은 그림 2와 같다.

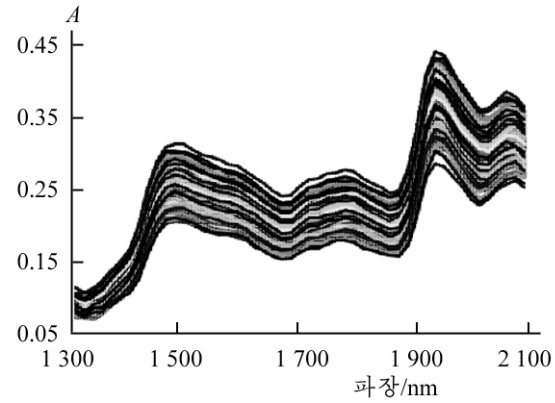

그림 2. 밀가루시료들의 근적외선흡수스펙트르

2) 실험자료준비

평등우연수발생원리에 기초하여 136개의 시료를 100개의 초기검량시료와 36개의 검정시료로 가르고 두 시료모임에 대하여 자료행렬을 구성한다. 이때 작성한 자료행렬의 크기는 각각 100×51, 36×51이고 자료행렬에 대응되는 측정값(단백질함량)벡토르의 크기는 각각 100×1, 36×1로 된다. 우와 같은 조작을 3번 반복하여 검량모임과 검정모임에서 시료들의 분포를 최대로 균일하게 한다.

3. 실험 및 결과해석

유전알고리듬을 리용하여 검량시료와 분석파장을 동시에 선택하는 경우와 검량시료를 먼저 선택하고 분석파장을 선택하는 경우, 분석파장을 먼저 선택하고 시료를 선택하는 경우 다중선형회귀(MLR), 주성분회귀(PCR), 부분최소두제곱법회귀(PLSR)모형에 의한 교차검정오차들을 계산하였다. 또한 련속사영알고리듬(SPA)을 리용하여 최적검량시료와 분석파장을 선택하고 우의 경우와 비교하였다.

초기시료모임을 주성분분석하여 얻은 제1주성분과 제2주성분을 자리표로 나타낸 최적검량시료들의 분포상태와 최적분석파장들의 분포상태는 각각 그림 3, 4와 같다.

그림 3에서 보는바와 같이 유전알고리듬과 련속사영알고리듬으로 다같이 선택한 검량시료들의 공간분포는 초기시료모임에 들어있는 100개 시료들의 분포와 비교적 잘 일치한다.

또한 그림 4에서 보는바와 같이 최적분석파장들이 측정파장구간에서 균등하게 분포되여있는것이 아니라 일부 구역들에 집중되여있다. 이것은 밀가루의 근적외선흡수스펙트르에서 단백질분자의 주요흡수봉우리들이 이 구역들에 존재한다는것을 의미한다.

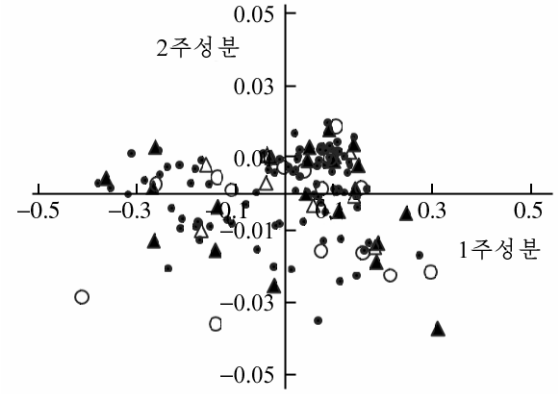

그림 3. 초기시료와 최적검량시료들의 분포상태
● 초기시료, ○ SPA로 선택한 검량시료, △ GA로 선택한 검량시료, ▲ SPA와 GA가 다같이 선택한 검량시료

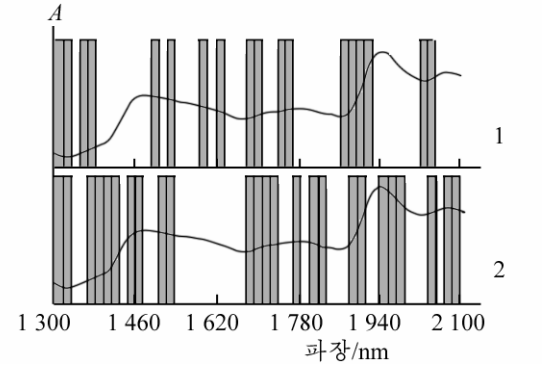

그림 4. 유전알고리듬(1)과 련속사영알고리듬(2)으로 선택한 최적분석파장들의 분포상태
주파장: 유전알고리듬의 경우에는 알고리듬을 실행할 때 매번 중복되여 선택되는 파장들을 의미, 련속사영알고리듬의 경우에는 직교성이 최대인 순서로 25개(선택된 파장들의 평균개수)의 파장들을 의미[4]

유전알고리듬에 의한 최적검량시료와 분석파장선택에서 평균2차뿌리예측오차는 표 2와 같다.

표 2. 유전알고리듬에 의한 최적검량시료와 분석파장선택에서 평균2차뿌리예측오차(RMSEP)

회귀방법	선택없음**	선택방법	파장선택*	시료선택*	시료－파장 선택**	파장－시료 선택**	동시선택**
MLR	1.772 (51/100)	GA	1.635(35)	1.324(45)	1.214(29/45)	1.043(35/42)	0.669(27/45)
		SPA	1.147(18)	1.575(63)	1.402(17/63)	1.037(18/86)	
PCR	1.287 (51/100)	GA	0.900(26)	0.679(46)	0.644(27/46)	0.655(26/43)	0.600(28/52)
		SPA	1.065(42)	0.998(72)	0.898(41/72)	0.941(42/65)	
PLSR	0.901 (51/100)	GA	0.835(24)	0.476(38)	0.470(32/38)	0.448(24/45)	0.412(28/39)
		SPA	0.901(51)	0.870(54)	0.870(45/54)	0.476(51/38)	

* 괄호안의 수값은 선택된 파장 혹은 시료개수, ** 괄호안의 수값은 파장개수/시료개수

표 2에서 보는바와 같이 유전알고리듬을 리용하여 검량시료와 분석파장을 동시에 선택하는 경우 시료와 파장을 개별적으로 선택하는 경우나 둘중 하나를 선택하고 다음것을 선택하는 경우에 비하여 훨씬 더 작은 예측오차를 가진 결과를 얻을수 있다.

맺 는 말

유전알고리듬을 리용하여 다변량검량모형작성을 위한 초기자료행렬에서 최적시료모임과 최적분석파장을 동시에 선택하는 방법을 확립하였다. 이 방법으로 밀가루시료의 근적외선흡수스펙트르자료에서 최적시료모임과 최적분석파장을 동시에 선택한 결과 개별적으로 선택하였을 때보다 더 좋은 예측결과를 주었다.

참 고 문 헌

[1] 서이식 등; MATLAB에 의한 유전알고리듬의 응용, 공업출판사, 4～10, 주체101(2012).

[2] A. Bogomolov et al.; Chemom. Intell. Lab. Syst., 126, 129, 2013.

[3] Howard Mark et al.; Chemometrics in Spectroscopy, Elsevier Inc., 256～315, 2007.

[4] H. A. D. Filho et al.; Chemom. Intell. Lab. Syst., 72, 83, 2004.

주체106(2017)년 4월 5일 원고접수

Simultaneous Selection of Optimal Calibration Samples and Wavelengths for Multivariate Calibration Modeling using the Genetic Algorithm

Choe Kang Jin, Pak Yong Gil and *Ri Ju Chil*

We proposed the simultaneous selection method of optimal calibration samples and wavelengths for multivariate calibration modeling using the genetic algorithm.

We simultaneously selected the optimal calibration samples and wavelengths from the NIR absorption spectrum of the wheat flour. As the result, it exhibits better prediction results than the individual selection method.

Key words: multivariate calibration modeling, simultaneous selection, genetic algorithms