다변량검량모형작성에서 유전알고리듬에 의한 시료모임분할방법

박영길, 최강진, 장영기

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《대학에서는 사회주의강국건설에서 나서는 리론실천적, 과학기술적문제들을 원만히 해결하며 기초과학부문을 발전시키고 첨단과학기술분야를 개척하는데 중심을 두고 과학연구사업을 진행하여야 합니다.》

다변량검량모형작성에서 주어진 시료들을 모형작성에 리용되는 검량시료모임과 작성한 모형의 평가에 리용되는 검정시료모임으로 나누는 시료모임분할은 모형의 예측정확성을 높이기 위한 관건적인 문제이다.

현재까지 시료모임분할방법으로는 우연시료채취법, 케나르드－스톤법[2], DUPLEX법[3] 등을 비롯한 여러가지 수법[4]들이 알려져있으나 분할하여 얻은 두 시료모임에서 시료들의 분포차이를 정확히 평가하기 힘든 부족점을 가지고있다.

우리는 주어진 시료들을 분포가 같은 두 부분모임으로 나눌수 있는 새로운 시료모임분할척도를 제기하고 최량탐색방법인 유전알고리듬과 결합하여 시료모임을 분할하는 방법을 새롭게 확립하였다.

1. 리론적기초

근적외선검량과정에는 흔히 선형모형을 리용한다. 따라서 이러한 모형작성에 리용되는 시료모임에서 시료들의 공간적분포는 우연량의 통계적분포와 류사한 방법으로 평가할수 있다. 이로부터 두 시료모임에서 시료분포가 될수록 일치하자면 다음과 같은 요구성을 최대로 만족하여야 한다.

첫째로, 매개 모임에서 시료들은 평등분포를 가져야 한다.

둘째로, 두 모임에서 시료들의 분포중심이 같아야 한다.

셋째로, 두 모임에서 시료들의 분산 또는 표준편차가 같아야 한다.

새로운 시료모임분할척도는 우의 요구성을 만족시키는 수학적인 함수들로 구성된다.

우선 매개 모임에서 시료들의 거리분포가 평등분포의 요구성에서 차이나는 정도와 관련된 함수를 제기하였다. 시료들을 다차원동심구들에 배치할 때 매개 시료들이 놓이게 될 다차원구의 반경은 식 (1)과 같이 표시할수 있다.

$$r_i = R \times \sqrt[N]{\frac{V_i}{V}} = R \times \sqrt[N]{\frac{i}{N}} \tag{1}$$

여기서 V-시료들이 분포된 공간의 총체적, V_i-주어진 공간의 체적을 등분하는 다차원동심구들의 체적, i-다차원동심구번호, N-주어진 공간차원수, R-주어진 전체 공간을 나타내는 다차원구의 반경이다.

이로부터 우리는 시료모임의 분포중심점을 자리표원점으로 하는 N차원자리표계에서 매 시료들을 자리벡토르크기순서로 배렬하였을 때 i번째 시료의 자리벡토르크기와 식 (1)로부터 계산된 구의 반경 r_i와의 차 d_i로부터 다음과 같은 함수를 제기하였다.

$$F_1(X)=\frac{\sqrt{\dfrac{\sum_{i=1}^{m}d_i^2}{m}}}{\overline{D_i}} \tag{2}$$

여기서 X는 분할하여 얻은 매개 시료모임에서 시료들의 스펙트르자료, $\overline{D_i}$는 시료자리벡토르들의 크기평균값이다.

이 함수는 정의 값을 가지며 주어진 시료들의 거리분포가 평등분포의 요구성에서 어긋나는 정도를 나타내는데 이 값이 클수록 시료들은 해당한 부분공간으로부터 멀리 떨어져있게 되며 이때 시료들의 거리분포가 평등분포에서 많이 차이나게 된다.

다음으로 모임에 속한 시료들의 자리벡토르방향이 평등분포의 요구성에서 차이나는 정도와 관련된 함수를 제기하였다. 매개 시료자리벡토르들사이 각을 계산하고 해당 시료자리벡토르가 가지는 최소각 $\cos\theta_{i,\min}$을 찾을 때 이 각들의 크기의 합이 작을수록 주어진 시료자리벡토르들은 어느 한 축주위에 혹은 좁은 방향에 분포되여있다고 평가할수 있다. 임의의 두 벡토르 $\boldsymbol{x}_i$, $\boldsymbol{x}_j$사이의 각 $\theta_{i,j}$는 다음의 식에 의하여 계산된다.

$$\cos\theta_{i,j}=\frac{\boldsymbol{x}_i\cdot\boldsymbol{x}_j}{|\boldsymbol{x}_i||\boldsymbol{x}_j|}=\frac{\sum_{k=1}^{N}\boldsymbol{x}_{ik}\boldsymbol{x}_{jk}}{\sqrt{\sum_{k=1}^{N}\boldsymbol{x}_{ik}^2}\sqrt{\sum_{k=1}^{N}\boldsymbol{x}_{jk}^2}} \tag{3}$$

$\cos\theta_{i,\min}$은 i번째 시료벡토르와 다른 모든 벡토르들사이 각의 코시누스값가운데서 최소값을 취한다. 이로부터 다음과 같은 함수를 설계하였다.

$$F_2(X)=\frac{\sum_{i=1}^{m}(\cos\theta_{i,\min}+1)}{m} \tag{4}$$

이 함수는 0부터 1사이의 값을 가지며 함수값이 클수록 자리벡토르들은 공간의 어느 한 방향으로만 놓일 가능성이 크다. 자리벡토르크기분포가 같은 두 시료모임에서도 이 함수값이 작은 모임일수록 시료들이 공간의 각이한 방향으로 보다 균일하게 분포될것이다.

다음으로 분할하여 얻은 두 부분시료모임에서 시료들의 분포중심과 표준편차가 각각 같아야 한다는 요구성들과 관련된 함수를 제기하였다.

매개 시료모임에서 시료들의 공간적분포는 분포중심을 구의 중심으로 하고 표준편차를 반경으로 하는 다차원구형태로 나타낼수 있다. 따라서 우리는 두 구의 겹침정도와 관련된 다음과 같은 함수를 구성하였다.

$$F_3(\boldsymbol{X}_1,\ \boldsymbol{X}_2)=1-\frac{\mathrm{std}(\boldsymbol{X}_1)+\mathrm{std}(\boldsymbol{X}_2)-d(\boldsymbol{X}_1,\ \boldsymbol{X}_2)}{2\times\max(\mathrm{std}(\boldsymbol{X}_1),\ \mathrm{std}(\boldsymbol{X}_2))} \tag{5}$$

여기서 $\boldsymbol{X}_1$, $\boldsymbol{X}_2$는 두 부분시료모임(검량 및 검정시료모임)의 스펙트르자료행렬, $\mathrm{std}(\boldsymbol{X}_1)$, $\mathrm{std}(\boldsymbol{X}_2)$는 두 부분시료모임에서 시료벡토르크기들의 표준편차, $\max(\mathrm{std}(\boldsymbol{X}_1),\ \mathrm{std}(\boldsymbol{X}_2))$는

두 표준편차가운데서 큰 값을 취하는 함수, $d(\boldsymbol{X}_1, \boldsymbol{X}_2)$ 는 자료행렬 $\boldsymbol{X}_1$의 평균값벡토르와 자료행렬 $\boldsymbol{X}_2$의 평균값벡토르사이 거리(두 분포중심사이의 거리)이다.

이 함수의 특성을 보면 두 모임에 속한 시료들의 분포중심이 일치하고 표준편차가 같을 때 함수값이 0으로 다가가며 중심점들의 거리가 멀고 두 표준편차의 차이가 커질수록 그 값은 증가한다. 또한 중심점들사이 거리가 같은 경우에도 표준편차들의 크기가 다 같이 커지는 경우 두 분포를 나타내는 구들의 겹치는 공간이 더욱 커지며 함수값도 증가한다.

3개의 함수를 결합시켜 적합한 분할척도로 되는 함수를 구성하는데서 기본요구는 함수식의 값변화와 분할된 두 부분모임들에서의 분포차이변화가 대응되게 하는것이다.

매 함수식이 가지는 값범위는 비교적 류사하지만 그 변화특성이 차이나므로 우리는 식 (2)와 (4), (5)에 각이한 무게를 주어 선형결합한 새로운 시료모임분할척도 $F(\boldsymbol{X}_1, \boldsymbol{X}_2)$를 구성하였다.

$$F(\boldsymbol{X}_1, \boldsymbol{X}_2) = w_1 \times [F_1(\boldsymbol{X}_1) + F_1(\boldsymbol{X}_2)] + w_2 \times [F_2(\boldsymbol{X}_1) + F_2(\boldsymbol{X}_2)] + w_3 \times F_3(\boldsymbol{X}_1, \boldsymbol{X}_2) \quad (6)$$

여기서 w_1, w_2, w_3은 무게곁수이다.

이 함수는 정의 값을 가지며 분할하여 얻은 두 시료모임에서 시료들의 분포가 평등분포의 요구성을 최대로 만족하고 시료들의 분포중심과 표준편차가 각각 일치할 때에는 최소값을 가지며 두 시료분포가 우에서 제기한 3가지 요구성에서 어긋날수록 보다 큰 값을 가진다.

또한 적합한 무게곁수를 선정하면 이 함수의 값변화는 두 부분모임들에서 분포차이변화를 반영하게 된다.

2. 시료모임분할을 위한 유전알고리듬파라메터설정

우리는 선행연구방법[1]에 따라 주목하는 분석물질인 성분 1을 포함한 3개의 성분과 비선형배경신호를 모방하여 만들었다. 이때 얻은 시료모임은 100개의 검량 및 검정시료와 100개의 예측시료들로 되여있다.

시료선택을 위한 알고리듬의 파라메터설정은 다음과 같이 하였다.[1]

적응도함수: 식 (6)을 그대로 리용

유전부호화: 2진부호화(매 염색체에 할당되는 변수개수 2)

군체크기: 50

회귀방법: 다중선형회귀, 주성분회귀, 부분최소2제곱회귀

교잡확률: 0.8

정지기준: 최대세대수 500, 무효세대수 250

3. 실험 및 결과해석

분할하여 얻은 두 부분시료모임에서 시료들의 분포차이를 정량적으로 평가하기 위하여 두 부분시료모임을 각각 검량모임으로 하여 다중선형회귀, 주성분회귀, 부분최소2제곱

회귀모형을 작성한다. 다음 얻은 모형을 리용하여 예측시료들에 대한 함량을 계산하여 오차를 비교하였다.(표 1)

표 1. 두 부분시료모임으로 각각 작성한 모형의 평균2차뿌리예측오차

부분시료모임번호	시료개수	MLR	PCR	PLSR
1	65	0.066 9	0.025 5	0.004 6
2	35	0.061 2	0.026 2	0.004 8

MLR－다중선형회귀, PCR－주성분회귀, PLSR－부분최소2제곱회귀

표 1에서 보는바와 같이 유전알고리듬을 리용하여 시료모임을 분할하는 경우 얻어진 두 부분모임을 각각 검량시료모임으로 하여 작성한 모든 회귀모형들의 예측정확성은 매우 류사하다. 이것은 두 시료모임에서 시료들의 분포가 거의 같다는것을 보여준다.

다음 유전알고리듬에 의한 시료분할방법을 종전의 방법들과 비교하기 위하여 우연시료채취법, 케나르드－스톤법, DUPLEX법, 유전알고리듬법을 리용하여 초기시료모임을 분할하였다. 이때 얻은 부분시료모임 1을 검량시료모임으로 하여 다중선형회귀, 주성분회귀, 부분최소2제곱회귀모형을 작성하고 이것을 리용하여 예측시료들에 대한 분석성분을 예측하여 평균2차뿌리예측오차를 계산하였다.(표 2)

표 2. 시료모임분할방법에 따르는 모형의 예측결과(평균2차뿌리예측오차)

분할방법	부분시료모임번호	시료개수	MLR	PCR	PLSR
분할하지 않음		100	0.099 6	0.026 4	0.004 8
우연시료채취법	1	65	0.134 0	0.027 2	0.007 1
	2	35	0.221 0	0.070 5	0.008 4
케나르드－스톤법	1	73	0.102 0	0.026 2	0.004 8
	2	27	0.184 0	0.087 1	0.007 4
DUPLEX법	1	71	0.086 4	0.024 3	0.004 6
	2	29	0.188 0	0.027 6	0.007 6
유전알고리듬법	1	65	0.066 9	0.025 5	0.004 7
	2	35	0.061 2	0.026 2	0.004 7

표 2에서 보는바와 같이 유전알고리듬법으로 시료를 분할하는 경우 두 부분시료모임을 각각 검량시료모임으로 리용하여 작성한 모형들의 예측오차가 제일 류사하면서도 우연시료채취법, 케나르드－스톤법, DUPLEX법들을 리용할 때보다 평균 1.5배 작은 오차값을 준다는것을 알수 있다.

맺 는 말

우리가 제기한 시료모임분할방법은 종전의 방법들보다 분할하여 얻은 두 시료모임의 분포일치성이 좋을뿐아니라 모형의 예측률이 높다.

참 고 문 헌

[1] A. Kalinin et al.; Talanta, **115**, 755, 2013.
[2] R. W. Kennard et al.; Technometrics, **11**, 37, 1969.
[3] R. D. Snee; Technometrics, **19**, 415, 1977.
[4] 詹雪艳 等; 光谱学与光谱分析, **34**, 2367, 2014.

주체107(2018)년 7월 5일 원고접수

Partition Method of Sample Set for Multivariate Calibration Modeling by Using Genetic Algorithm

Pak Yong Gil, *Choe Kang Jin* and *Jang Yong Gi*

We proposed a new partition method of sample set for multivariate calibration modeling by using genetic algorithm.

In this method the agreement of the distribution of partitioned two sample set is better and also the prediction ability of the model is higher than previous partition methods such as random sampling method, the Kennard–Stone method and the DUPLEX method.

Key words: sample set partition, genetic algorithm