근사모델의 분산과 신뢰구간을 이용한 모델의 정확도 평가법

한인식† · 이용빈* · 최동훈**

Validation Technique using variance and confidence interval of metamodel

Insik Han, Yongbin Lee and Dong-Hoon Choi

Key Words : Accuracy(정확도), Metamodel(근사모델), Validation technique(정확도 평가법), Varianve(분산), Confidence interval(신뢰구간)

Abstract

The validation technique is classified with two methods whether to demand of additional experimental points. The method which requires additional experimental points such as RSME is actually impossible in engineering field. Therefore, the method which only use experimented points such as the cross validation technique is only available. But the cross validation not only requires considerable computational costs for generating metamodel each iterations, but also cannot measure quantitatively the fidelity of metamodel. In this research we propose a new validation technique for representative metamodels using an variance of metamodel and confidence interval information. The proposed validation technique computes confidence intervals using a variance information from the metamodel. This technique will have influence on choosing the accurate metamodel, constructing ensemble of each metamodels and advancing effectively sequential sampling technique.

1. 서 론

현재의 공학분야에서는 실험에 소요되는 시간과 비용을 줄이기 위해 컴퓨터를 이용한 시뮬레이션 모델을 많이 사용하고 있다. 하지만 이러한 컴퓨터 시뮬레이션 모델도 실제 실험과의 오차를 줄이기 위해 점점 복잡해 지고 있으며, 따라서 시뮬레이션 모델을 사용하는 것도 비용이 많이 들기는 마찬가지이다. 이러한 문제점을 극복하기 위하여 복잡한 시스템 모델을 대신할 수 있는 수학적인 근사모델(Approximate model or Surrogate model or Metamodel)을 활용하는 많은 연구가 진행되고 있다. 이러한 근사모델 기법 중 현재 많이 사용되고 있는 근사모델로는 일반적으로 반응표면 모델(RSM: Response Surface Model)이라고 불리는 다항 회귀 모델(Polynomial Regression), 크리깅(Kriging), 방사 기저 함수(Radial Basis Functions), 서포트 벡터 회귀 기법(Support Regression Vector)등이 있다.

그러나 대부분의 공학 시스템의 경우 실제모델이 어떤 경향을 갖는지 미리 알지 못하는 상태이기 때문에 어떠한 근사모델이 실제모델을 정확하게 표현했는지 알 수 없다. 따라서 생성된 근사모델의 정확도를 평가하는 기법에 대한 연구는 근사모델 기반 최적설계의 정확성을 보장하기 위해 반드시 필요하다. 근사모델의 정확도 평가 기법은 추가 실험점의 요구 여부로 크게 두 가지로 분류할 수 있다. 먼저 추가 실험점을 요구하는 기법(e.g., RMSE, Maximum Error, Average Error) 은 값비싼 해석을 추가로 요구하기 때문에 실제 공학 문제에 적용하는 것은 현실적으로 불가능 하다. 따라서 추가 실험점이 필요없는 방법을 사용해야 하며, 주로 사용되고 있는 기법으로는 교차검증법(cross-validation)이 있다[1]. 교차검증법은 교차검증 오차를 계산하기 위해 여러 번의 모델을 생성하는 것을 요구하는 단점이 있다. 더욱이 교차검증법은 정량적으로 근사모델의 정확성을 나타내주지 못하며 오직 생성한 근사모델이 실험점에 얼마나 민감한 가를 보여 줄 뿐이다.

† 한양대학교 대학원 기계공학과
E-mail : ishan83@gmail.com
TEL : (02)2220-0478 FAX : (02)2290-4070
* 한양대학교 대학원 기계공학과
** 회원, 최적설계신기술연구센터 소장, 한양대학교 기계공학부 교수

본 연구에서는 분산과 신뢰구간 정보를 이용하여 대표적인 근사모델의 정확도를 평가하는 평가기법을 제안하고자 한다. 이 기법은 근사모델의 분산정보를 얻어내어서 실험점의 신뢰구간을 계산하여 근사모델의 정확도를 평가하는 기법이다. 여기서 얻어진 근사모델의 분산정보는 설계변수에 대한 함수로 표현되기 때문에, 기존 기법들처럼 전체 설계영역에서의 평균적인 근사모델의 정확성을 평가하는 것이 아니라 각 예측점에서 근사모델의 정확도를 평가 할 수 있다는 장점이 있다. 본 연구의 파급효과로 정확한 근사모델의 선택 가능, 적합한 근사모델의 앙상블(ensemble of metamodels) 구현 및 근사모델을 이용한 효과적인 순차적 추출법(Sequential Sampling)의 발전 등을 기대한다.

2. 근사모델 기법

2.1 다항 회귀법 (Polynomial Regression)

다항 회귀 모델은 설계 영역 안에 있는 실험점들을 최소제곱법(Least Square Method, LSM) 을 이용하여 근사화 하는 방법이다[2]. 공학분야에서 보통 많이 사용되는 2 차 다항식 형태의 모델은 다음과 같다.

$$\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \beta_i x_i^2 + \sum_{i=1}^{n} \sum_{j=1, i<j}^{n} \beta_{ij} x_i x_j \quad (1)$$

여기서 n 은 설계 변수의 개수, x_i 는 설계변수의 값을, $\hat{y}$ 은 근사모델로부터 구한 예측 값을 의미하고, β_i 와 β_{ij} 는 회귀계수이다. β 는 예측된 값들의 편차 제곱의 합을 최소로 하는 최소제곱법을 이용하여 계산한다

다항 회귀모델은 생성이 용이하고, 노이즈를 완화시키는 효과가 있으므로 수치적 노이즈가 있는 모델에 유용하게 사용될 수 있다. 하지만, 이러한 저차(low order)의 함수로는 복잡한 형태의 비선형함수를 근사화하는 데에 적절하지 않으며, 고차의 반응표면모델을 사용한다 하더라도 불안정한 근사함수를 생성할 수 있으며 고차로 갈수록 필요한 실험점의 수가 많아진다는 단점이 있다. 또한 공학 분야에서 많이 사용되고 있는 2 차 회귀 모델을 사용할 경우 설계변수가 많아지면, 필요한 실험점의 개수가 기하급수적으로 증가하여 효율성이 급격히 떨어진다는 단점이 있다.

2.2 서포트 벡터 회귀 기법 (Support Vector Regression, SVR)

서포트 벡터 회귀 기법[3]에서의 선형 회귀 모델의 형태는 다음과 같이 나타낸다.

$$f(x) = \langle w \cdot x \rangle + b \quad (2)$$

여기서 $\langle w \cdot x \rangle$ 는 w 와 x 의 내적을 의미한다. 모델과 인접한 점(support vector)과의 거리는 $1/\|\mathbf{w}\|$ 로 나타내어 정해진 범위 ε 내에서 $\|\mathbf{w}\|^2$ 을 최소화 하는 식 (3)의 최적화 문제를 통해 최종 모델을 결정한다.

$$\begin{aligned} &\textit{Minimize} \quad \frac{1}{2}|w|^2 \\ &\textit{subject to} \begin{cases} y_i - \langle w \cdot x_i \rangle - b \le \varepsilon \\ \langle w \cdot x_i \rangle + b - y_i \le \varepsilon \end{cases} \end{aligned} \quad (3)$$

라그랑지 이론을 통해 원래 최적화 문제인 식 (3)을 쌍대(dual) 문제 형태로 바꾸면 식 (4)이 되며,

$$\begin{aligned} \textit{Maximize} \quad & -\frac{1}{2}\sum_{i,j=1}^{nexp}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle \\ & -\varepsilon\sum_{i=1}^{nexp}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{nexp} y_i(\alpha_i - \alpha_i^*) \\ \textit{subject to} \quad & \sum_{i=1}^{nexp}(\alpha_i - \alpha_i^*) = 0 \\ & \alpha_i \,,\, \alpha_i^* \in [0, C] \end{aligned} \quad (4)$$

최종적으로 얻게 되는 회귀 모델은 식 (5)과 같이 나타낼 수 있다.

$$f(\mathbf{x}) = \sum_{i=1}^{nexp}(\alpha_i^* - \alpha_i)\langle \mathbf{x}_i \cdot \mathbf{x} \rangle + b. \quad (5)$$

서포트 벡터 회귀 기법에는 다음과 같은 커널 함수들이 있다.

Table 1 Kernel functions

Linear	$k(\mathbf{x},\mathbf{x}') = \mathbf{x}^{\mathrm{T}}\mathbf{x}'$
Polynomial	$k(\mathbf{x},\mathbf{x}') = \langle \mathbf{x} \cdot \mathbf{x}' \rangle^d$
Gaussian	$k(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$
Sigmoid	$k(\mathbf{x},\mathbf{x}') = \tanh\left(\kappa\langle \mathbf{x} \cdot \mathbf{x}' \rangle + \vartheta\right)$
Inhomogeneous Polynomial	$k(\mathbf{x},\mathbf{x}') = \left(\langle \mathbf{x} \cdot \mathbf{x}' \rangle + c\right)^d$

Table 1 에서 보여진 5 개의 kernel 함수 중 하나를 선택하여 식 (4)의 내적 항 대신 사용하면 비선형 회귀 모델을 구할 수 있다.

$$Maximize \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{nexp}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)k\langle \mathbf{x}_i\cdot\mathbf{x}_j\rangle \\ -\varepsilon\sum_{i=1}^{nexp}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{nexp}y_i(\alpha_i-\alpha_i^*) \end{cases}$$

$$subject\ to \begin{cases} \sum_{i,j=1}^{nexp}(\alpha_i-\alpha_i^*)=0 \\ \alpha_i\ ,\ \alpha_i^*\in[0,C] \end{cases} \quad (6)$$

식 (6)은 식 (4)의 내적 부분을 커널 함수 식으로 대체한 경우 이며, 이 식을 통해 비선형 회귀모델을 구하면 식 (7)과 같이 나타낼 수 있다.

$$f(x)=\sum_{i=1}^{nexp}(\alpha_i^*-\alpha_i)k\langle \mathbf{x}_i\cdot\mathbf{x}\rangle+b \quad (7)$$

이 방법은 실험점의 개수에 대한 제약이 없고, 회귀 모델을 만드는데 실험점의 개수만큼의 성분을 갖는 벡터를 내적하는 비교적 간단한 계산으로 근사모델을 얻을 수 있다는 장점이 있다.

2.3 방사기저함수 (Radial Basis Function, RBF)

RBF[4]는 산재된 다 변량의 데이터 보간을 위해 제안되었다. 근사 모델의 형태는 유클리드 거리(Euclidean distance)와 강도(weight)의 선형조합으로서 식 (8)과 같이 나타낸다

$$\hat{y}(\mathbf{x})=\sum_{i=1}^{nexp}w_i\phi_i(\mathbf{x},x_i) \quad (8)$$

이때 *nexp* 는 실험점의 개수, w_i는 최소자승법에 의해 결정된 강도를 의미하고, $\phi_i(\mathbf{x},x_i)$는 실험점 x_i에 의해 정의된 i 번째 기저함수이다. 여기서 기저함수는 Table 2 에 나타낸 다양한 대칭의 방사형 함수들이 사용된다.

이러한 특징을 갖는 RBF 방법은 Kriging 과 함께 아주 좋은 성능을 나타내는 보간 기법으로 알려져 있다. 하지만, 이러한 보간 기법은 기저함수의 형태와 기저함수에 내에 존재하는 파라미터의 값에 따라 모델의 형상이 많이 달라진다는 단점을 가지고 있다.

2.4 크리깅 (Kriging)

크리깅 모델[5]은 전산실험으로 얻은 실험점의 정보를 전역모델과 국부편차의 합으로 표현하며, 다음 식과 같이 가정한다.

$$y(x)=\mathbf{f}(\mathbf{x})\boldsymbol{\beta}+Z(\mathbf{x}) \quad (9)$$

여기서 $Z(\mathbf{x})$는 평균이 0 이고 분산이 σ^2인 정규분포를 따르고 각 실험점에서의 편차들은

Table 2 Radial Functions for RBF Model

Name	Radial Function
Gaussian	$h(\mathbf{x})=\exp\left(-\frac{(\mathbf{x}-\mathbf{c})^T(\mathbf{x}-\mathbf{c})}{r^2}\right)$
multiquadric	$h(\mathbf{x})=\sqrt{1+\frac{(\mathbf{x}-\mathbf{c})^T(\mathbf{x}-\mathbf{c})}{r^2}}$
inverse multiquadric	$h(\mathbf{x})=\frac{1}{\sqrt{1+\frac{(\mathbf{x}-\mathbf{c})^T(\mathbf{x}-\mathbf{c})}{r^2}}}$
Cauchy	$h(\mathbf{x})=\frac{1}{1+\frac{(\mathbf{x}-\mathbf{c})^T(\mathbf{x}-\mathbf{c})}{r^2}}$

상관관계를 가지며 식 (10)으로 정의할 수 있다.

$$Cov\left[Z(\mathbf{x}^i),Z(\mathbf{x}^j)\right]=\sigma^2\mathbf{R}\left[R(\mathbf{x}^i,\mathbf{x}^j)\right],\quad i,j=1,...,nexp$$
$$R(\mathbf{x}^i,\mathbf{x}^j,\theta)=\exp\left[-\sum_{k=1}^{n}\theta_k\left|x_k{}^i-x_k{}^j\right|^2\right] \quad (10)$$

이 때 *nexp* 와 *n* 은 각각 실험점의 개수와 설계변수의 개수를 나타내며, $x_k{}^i$ 는 i 번째 실험점의 k 번째 설계변수의 값을 의미한다. 또한 $R(\mathbf{x}^i,\mathbf{x}^j)$는 임의의 두 실험점 $\mathbf{x}^i$와 $\mathbf{x}^j$의 상관관계를 표현한 함수이며 주로 가우시안 상관함수(Gaussian correlation function)가 사용된다. 최종적으로 크리깅 모델을 구성하기 위해서는 최우량추정법(Maximum Likelihood Estimation , MLE)을 통해 상관인자 θ_k를 결정한다. 한편 식 (9)의 크리깅 모델은 다음과 같이 유도된다

$$\hat{y}(\mathbf{x})=\mathbf{f}(\mathbf{x})^T\hat{\boldsymbol{\beta}}+\mathbf{r}(\mathbf{x})^T\gamma^* \quad (11)$$

여기서 $\hat{\boldsymbol{\beta}}=(\mathbf{F}^T\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{R}^{-1}\mathbf{Y}$, $\gamma^*=\mathbf{R}^{-1}(\mathbf{Y}-\mathbf{F}\hat{\boldsymbol{\beta}})$로 표현된다. $\mathbf{r}(\mathbf{x})$는 예측점과 실험점들간의 상관관계를 나타내는 상관벡터로 식(10)을 이용하여 나타내면 식 (12)와 같이 나타낼 수 있다.

$$\mathbf{r}(\mathbf{x})=[\mathbf{R}(\mathbf{x},\mathbf{x}^1),\mathbf{R}(\mathbf{x},\mathbf{x}^2),\mathbf{L},\mathbf{R}(\mathbf{x},\mathbf{x}^n)]^T \quad (12)$$

크리깅 모델이 구해지는 과정에서 평균제곱오차(Mean Squared Error, MSE)는 다음과 같은 식으로 표현된다

$$MSE=\sigma^2\left(1-\begin{bmatrix}\mathbf{f}(\mathbf{x})^T & \mathbf{r}(\mathbf{x})^T\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{F}^T\\ \mathbf{F} & \mathbf{R}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{f}(\mathbf{x})\\ \mathbf{r}(\mathbf{x})\end{bmatrix}\right) \quad (13)$$

식 (10)과 식(13)의 모수인 분산은 *nexp* 개의 데

이터들로 설명되지 않는 오차에 대한 추정된 분산 (Estimated variance)이며 식(14)과 같다.

$$\sigma^2 = \frac{(\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}})}{nexp} \quad (14)$$

크리깅 모델은 상관인자 θ_k 를 결정하는 과정에서 전역 최적설계 과정을 통해 이루어 지기 때문에 설계변수가 많고 실험점이 다수 존재하는 문제에 있어서 큰 계산비용을 요구한다는 단점을 가지고 있다. 또한 전역 최적설계가 제대로 이루어지지 않을 경우 근사모델을 올바르게 생성하지 못한다는 단점이 있다.

3. 기존의 정확도 평가법

3.1 추가실험이 필요한 정확도 평가법

추가실험이 필요한 정확도 평가법은 추가 실험점을 선택하여 근사모델의 예측값과 시뮬레이션의 참값과의 차이를 비교해서 오차를 평가하는 방법이다. 이러한 방법에는 평균제곱근오차(Rood Mean Squared Error, RMSE), 최대오차(Maximum Error) 그리고 평균오차(Average Error)등이 있으며, 각 방법에 대한 식은 다음과 같다.

$$E_{RMS} = \sqrt{\frac{1}{n_{v-add}} \sum_{i=1}^{n_{v-add}} (y_i - \hat{y}_i)^2} \quad , i = 1, \cdots, n_{v-add}$$

$$E_{MAX} = \max \left| y_i - \hat{y}_i \right| \quad , i = 1, \cdots, n_{v-add} \quad (15)$$

$$E_{AV} = \frac{1}{n_{v-add}} \sum_{i=1}^{n_{v-add}} \left| y_i - \hat{y}_i \right|^2 \quad , i = 1, \cdots, n_{v-add} \quad (16)$$

$$(17)$$

여기서 $n_{v\text{-}add}$ 는 추가실험점의 개수를 의미한다. 평균제곱근오차와 평균오차는 설계영역 전체 근사모델의 정확도를 평가하기 위해 사용되는 정확도 평가법이다. 반면에 최대오차는 근사모델의 국부적인 변화의 크기를 평가하는 기준이다.

이러한 평가법은 보간모델의 정확도 사용될 수는 있지만, 정확한 평가를 위해서 상당히 많은 추가실험점이 필요하다. 이것은 실제 공학 문제에서 상당한 수치적 부담을 야기시킬 수 있는 단점이 있다.

3.2 추가실험이 필요 없는 정확도 평가법

현실적으로 적용 가능한 검증기법이 요구됨에 따라 k 점 선택 교차검증법이라는 정확도 평가법이 제안되었다[6]. 교차검증법은 근사모델을 구성하기 위해 선택된 전체 실험점의 개수 $nexp$ 를 근사모델을 재구성하기 위한 n_c 와 근사모델의 오차를 평가하기 위한 검증점 n_v 로 나누고 실험점 n_c 을 이용하여 근사모델을 재구성한 후 실험점 n_v 에 대하여 오차를 평가하는 방법이다. 이 때, 오차를 평가하기 위한 검증점 n_v 의 개수는 k 와 같은 값이면 오차를 평가하기 위해 전체 실험점에서 k 점만큼 선택해야 하기 때문에 k 점 선택 교차검증법이라 한다. 이중 많이 쓰이고 있는 1 점 선택 교차검증법의 식은 다음과 같다.

$$CV = \sqrt{\frac{1}{nexp} \sum_{i=1}^{nexp} (\hat{Y}_{-i}(\mathbf{x}_i) - Y(\mathbf{x}_i))^2} \quad (18)$$

여기서 $Y(\mathbf{x}_i)$ 는 i 번째 실험점에서의 실제 응답값을 의미하고, $\hat{Y}_{-i}(\mathbf{x}_i)$ 는 i 번째 실험점만 제외하고 모델을 구성하여 구한 예측값이다.

이 방법의 큰 장점은 검증을 위한 별도의 실험점을 선택하지 않고 이미 사용된 실험점 중에서 일부를 이용하여 근사모델의 정확도를 측정할 수 있다는 것이다. 그러나 n_v 를 어떻게 정하느냐에 따라 다양한 조합의 수가 존재하며 n_v 의 선택이 커질수록 정확도 검증을 위한 계산비용은 현저하게 증가한다. 또한 근사화가 정확하게 된 모델임에도 불구하고 검증을 위한 실험점에 대해 근사모델의 민감도가 클 경우 부정확한 모델로 평가될 수 있다는 단점이 있다. 결과적으로 교차검증법은 근사모델이 얼마나 실제모델을 잘 표현 했나를 측정하는 척도라기 보다는 생성된 근사모델이 실험점에 얼마나 민감한가를 나타내는 척도라고 할 수 있기 때문에 근사모델의 정확도 평가기법으로 적합하지 않다.

4. 분산과 신뢰구간을 이용한 정확도 평가법

4.1 근사모델의 분산

4.1.1 다항 회귀법(Polynomial Regression)

1 차 다항 회귀법에 의한 실험값의 예측값은 다음과 같이 나타낼 수 있다.

$$\hat{y} = \beta_0 + \beta_1 x \quad (19)$$

설계변수로 이루어진 벡터 $\mathbf{x}$를 정의하면

$$\mathbf{x}^T = (1, x) \tag{20}$$

다음과 같이 쓸 수 있다.

$$\hat{y} = (1, x)\begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix} = \mathbf{x}^T\boldsymbol{\beta} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} \tag{21}$$

$\hat{y}$는 β_0와 β_1의 선형 조합이기 때문에 $\hat{y}$의 분산 $V(\hat{y})$는 다음과 같이 나타낼 수 있다[7].

$$V(\hat{y}) = V(\beta_0) + 2x\,\mathrm{cov}(\beta_0, \beta_1) + x^2 V(\beta_1) \tag{22}$$

이를 행렬의 형태로 표현할 수 있다.

$$\begin{aligned} V(\hat{y}) &= [1, x]\begin{bmatrix} V(\beta_0) & \mathrm{cov}(\beta_0, \beta_1) \\ \mathrm{cov}(\beta_0, \beta_1) & V(\beta_1)\end{bmatrix}\begin{bmatrix}1 \\ x\end{bmatrix} \\ &= \mathbf{x}^T(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2\mathbf{x} \\ &= \mathbf{x}^T(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}\sigma^2 \end{aligned} \tag{23}$$

여기서 $\mathbf{X}$ 행렬은 1 차 다항 회귀법에 의한 설계 행렬이며 σ^2는 모분산이다.

다항 회귀법에서 y는 평균이 0, 분산이 σ^2인 오차의 조합으로 표현 할 수 있다.

$$y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon, \quad \mathrm{E}(\varepsilon)=0, \quad \mathrm{V}(\varepsilon)=\sigma^2\mathbf{I} \tag{24}$$

따라서 다항 회귀 모델의 분산은 다음과 같다

$$\begin{aligned} V(y) &= V(\hat{y}) + V(\varepsilon) \\ &= \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}\sigma^2 + \sigma^2 \\ &= \sigma^2(\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x} + 1) \end{aligned} \tag{25}$$

식 (23),(24),(25)의 모분산은 근사모델이 실제 함수를 잘 표현한다는 가정하에 모추정분산 s^2 으로 쓸 수 있으며 다음과 같이 표현한다.

$$s^2 = \frac{\sum_i^{nexp}(y_i - \hat{y}_i)^2}{nexp - 1} \tag{26}$$

여기서 *nexp* 는 실험 횟수를 의미한다.

4.1.2 방사 기저 함수(Radial Basis Functions,RBF)

대표적인 기저함수인 가우시안(Gaussian) 함수를 이용하면 실험값의 예측값은 다음과 같이 나타낼 수 있다

$$\hat{y}(x) = \sum_{i=1}^{nexp} w_i h_i(x_i), \qquad h(x) = \exp\left(-\frac{(x-c)^2}{r^2}\right) \tag{27}$$

이를 바탕으로 설계 행렬 $\mathbf{H}$를 형성하여 나타내면 다음과 같다.

$$\hat{y} = \mathbf{Hw}, \quad \mathbf{H} = \begin{bmatrix} h_1(x_1) & h_2(x_1) & \cdots & h_{nexp}(x_1) \\ h_1(x_2) & \ddots & & h_{nexp}(x_2) \\ \vdots & & \ddots & \vdots \\ h_1(x_n) & h_2(x_n) & \cdots & h_{nexp}(x_n) \end{bmatrix} \tag{28}$$

여기서 n은 설계변수의 개수를 의미한다.

또한 최소자승법에 의해서 결정된 $\hat{\mathbf{w}}$는 다음과 같다.

$$\hat{\mathbf{w}} = \mathbf{A}^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{y}, \quad \mathbf{A}^{-1} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1} \tag{29}$$

실험하지 않은 점들 $\mathbf{x}$로 이루어진 새로운 벡터 $\mathbf{z}_0$를 정의하면 예측값은 다음과 같다

$$\hat{\mathrm{y}} = \mathbf{z}^{\mathrm{T}}\hat{\mathbf{w}}, \quad \mathbf{z} = \begin{Bmatrix} h(x_1) \\ h(x_2) \\ \vdots \\ h(x_n) \end{Bmatrix} \tag{30}$$

방사기저함수에서 y는 평균이 0, 분산이 σ^2인 오차의 조합으로 표현 할 수 있다.

$$y = \mathbf{z}^T\hat{\mathbf{w}} + \varepsilon, \quad \mathrm{E}(\varepsilon)=0, \quad \mathrm{V}(\varepsilon)=\sigma^2\mathbf{I} \tag{31}$$

따라서 방사기저함수의 분산은 다음과 같다

$$\begin{aligned} V(y) &= V(\hat{y}) + V(\varepsilon) \\ &= \mathbf{z}^T(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}\sigma^2 + \sigma^2 \\ &= \sigma^2(\mathbf{z}^T(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z} + 1) \end{aligned} \tag{32}$$

4.1.3 서포트 벡터 회귀 기법(Support Vector Regression,SVR)

대표적인 기저함수인 가우시안(Gaussian) 함수를 이용하면 실험값의 예측값은 다음과 같이 나타낼 수 있다.

$$\hat{y}(x) = \sum_{i=1}^{nexp}\left(\alpha_i^* - \alpha_i\right)h_i(x) + b,$$

$$h_i(x) = \exp\left(-\frac{\left\|\mathbf{x}_i - \mathbf{x}'\right\|^2}{2\sigma^2}\right) \tag{33}$$

이를 바탕으로 설계 행렬 $\mathbf{H}$를 형성하여 나타내면 다음과 같다.

$$\hat{y} = \mathbf{Hw} + \mathbf{b}, \quad \mathbf{H} = \begin{bmatrix} h_1(x_1) & h_2(x_1) & \cdots & h_{nexp}(x_1) \\ h_1(x_2) & \ddots & & h_{nexp}(x_2) \\ \vdots & & \ddots & \vdots \\ h_1(x_n) & h_2(x_n) & \cdots & h_{nexp}(x_n) \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} \left(\alpha_1^* - \alpha_1\right) \\ \vdots \\ \left(\alpha_{nexp}^* - \alpha_{nexp}\right) \end{bmatrix} \tag{34}$$

실험하지 않은 점들 $\mathbf{x}$로 이루어진 새로운 벡터 $\mathbf{z}_0$를 정의하면 예측값은 다음과 같다

$$\hat{y} = \mathbf{z}^{\mathbf{T}}\mathbf{w} + b, \quad \mathbf{z} = \begin{Bmatrix} h(x_1) \\ h(x_2) \\ \vdots \\ h(x_n) \end{Bmatrix} \tag{35}$$

서포트 벡터 회귀 기법에서 y 는 평균이 0, 분산이 σ^2 인 오차의 조합으로 표현 할 수 있다.

$$\begin{aligned} y &= \mathbf{z}^T\mathbf{w} + b + \varepsilon, \\ \mathrm{E}(\varepsilon)=0, &\quad \mathrm{V}(\varepsilon)=\sigma^2\mathbf{I} \end{aligned} \tag{36}$$

따라서 서포트 벡터 회귀모델의 분산은 다음과 같다

$$\begin{aligned} V(y) &= V(\hat{y}) + V(\varepsilon) \\ &= \mathbf{z}^T(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}\sigma^2 + \sigma^2 \\ &= \sigma^2(\mathbf{z}^T(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z} + 1) \end{aligned} \tag{37}$$

4.1.4 크리깅(Kriging)

크리깅 모델은 관측값들의 선형 조합으로 표현할 수 있는 선형 예측자이며 다음과 같다[8].

$$\hat{y}(\mathbf{x}) = \mathbf{c}(\mathbf{x})^{\mathbf{T}}\mathbf{y} \tag{38}$$

$\mathbf{c(x)}$ 는 $\mathbf{x}$ 의 함수로 이루어진 $n\times1$ 벡터이다. 가능한 모든 선형 예측자중에 가장 실제 함수와 가까운 관측자를 얻기 위해 평균제곱오차를 정량화하면 다음과 같다.

$$\begin{aligned} MSE[\hat{y}(\mathbf{x})] &= E[(\hat{y}(\mathbf{x}) - y(\mathbf{x}))^2] \\ &= \mathrm{var}[y(\mathbf{x})] + \mathrm{var}[\hat{y}(\mathbf{x})] \\ &\quad + (E[\hat{y}(\mathbf{x})] - E[\hat{y}(\mathbf{x})])^2 \\ &\quad - 2\,\mathrm{cov}[\hat{y}(\mathbf{x}), y(\mathbf{x})] \end{aligned} \tag{39}$$

여기에 불편치 조건(Unbiased Condition)을 삽입하여 다시 쓰면 다음과 같다.

$$\begin{aligned} MSE[\hat{y}(\mathbf{x})] &= \mathrm{var}[y(\mathbf{x})] + \mathrm{var}[\hat{y}(\mathbf{x})] \\ &\quad - 2\,\mathrm{cov}[\hat{y}(\mathbf{x}), y(\mathbf{x})] \end{aligned} \tag{40}$$

MSE 를 최소화 하는 $\mathbf{c(x)}$ 는 KKT 필요조건(Karusch-Kuhn Tucker necessary conditions)에 의해 구해지며 식 (41)과 같이 나타낸다

$$\mathbf{c(x)} = \mathbf{R^{-1}[r(x) - F(F^TR^{-1}F)^{-1}(F^TR^{-1}r(x) - f(x))]} \tag{41}$$

따라서 $Y(\mathbf{x})$ 의 분산 $V[Y(\mathbf{x})]$ 는 식 (42)와 같이 나타낼 수 있다.

$$\begin{aligned} \mathrm{var}[y(\mathbf{x})] &= \mathrm{var}[\hat{y}(\mathbf{x}) + \varepsilon] \\ &= E[\hat{y}(\mathbf{x}) - E[\hat{y}(\mathbf{x})]^2] + \sigma^2 \\ &= E[\mathbf{c(x)}^{\mathbf{T}}\mathbf{ZZ}^{\mathbf{T}}\mathbf{c(x)}] + \sigma^2 \\ &= \sigma^2[\mathbf{c(x)}^{\mathbf{T}}\mathbf{Rc(x)} + 1] \end{aligned} \tag{42}$$

4.2 각 근사모델의 분산정보를 이용한 신뢰구간의 계산

실제 모델의 값과 그 근사모델이 제공하는 예측값이 일치하는 경우는 흔치 않다. 따라서 실제모델의 값이 포함되리라고 예측되는 구간을 추정하는 것을 구간추정(interval estimation)이라고 한다. 이 구간에 실제모델의 값이 포함될 확률을 신뢰수준(confidence level) 및 신뢰계수(confidence coefficient)라 하며, 이 구간을 신뢰구간(confidence interval)이라 한다. 또한 이 구간의 상한과 하한을 신뢰한계(confidence limit)라 한다. 신뢰구간을 식으로 나타내면 다음과 같다.

$$y(\mathbf{x}) \in \left[\, \hat{y}(\mathbf{x}) - z_a\sigma_y, \hat{y}(\mathbf{x}) + z_a\sigma_y \,\right] \tag{43}$$

여기서 $\mathbf{Z}_\alpha$ 는 신뢰수준이며 σ_y 는 근사모델의 표준편차이다. 근사모델의 표준편차는 분산의 제곱근 $\sqrt{V(y(\mathbf{x}))}$ 으로 계산 가능하다. 따라서 근사모델의 분산을 이용하여 예측값의 신뢰 구간을 계산할 수 있다. 예를 들어 95%의 신뢰수준을 바탕으로 한 예측값의 신뢰구간은 다음과 같다.

$$y(\mathbf{x}) \in [\hat{y}(\mathbf{x}) - 1.96\sqrt{V[y(\mathbf{x})]}, \hat{y}(\mathbf{x}) + 1.96\sqrt{V[y(\mathbf{x})]}] \tag{44}$$

본 연구에서 제안한 기법은 신뢰구간을 이용하여 근사모델의 정확도를 평가한다. 즉, 예측값에 대한 신뢰구간이 좁으면 좁을수록 근사모델의 정확도가 높다고 판단할 수 있다. 또한 신뢰구간이 설계변수의 함수로 표현되기 때문에 원하는 예측값에 대한 신뢰구간을 측정하여 정확도를 평가할 수 있다는 장점이 있다.

5. 결론 및 향후 연구

본 연구에서는 분산과 신뢰구간 정보를 이용하여 공학분야에서 많이 사용되고 있는 다항 회귀법(Polynomial Regression), 서포트 벡터 회귀 기법(Support Vector Regression, SVR), 방사 기저함수(Radial Basis Functions, RBF), 크리깅(Kriging) 근사모델들의 정확도 평가 기법을 제안하였다. 기존의 정확도 평가 기법과는 달리 추가 실험점이 필요 없을 뿐만 아니라 수치적인 방법이 아닌 수학적인 식으로 분산을 계산하기 때문에 큰 비용 없이 정확도를 평가 할 수 있는 것이 장점이다. 또한 제안한 기법으로 얻어진 근사모델의 분산정보는 설계변수에 대한 함수로 표현되기 때문에, 기존의

기법들처럼 전체 설계영역에서의 평균적인 근사모델의 정확도를 평가하는 것이 아니라 각 예측점에서 근사모델의 정확도를 평가 할 수 있다는 장점이 있다.

향후 다양한 예제를 통해 제안된 기법의 검증을 할 계획이며, 각 문제에서 실제 모델을 가장 잘 표현하는 근사모델을 선택하는 방법에 대한 연구와 여러 근사모델들의 앙상블 구현 및 순차적 근사 최적설계를 위한 순차적 추출법에 대한 연구를 실행할 예정이다.

후 기

본 연구는 최적설계신기술연구센터와 두뇌한국21 사업에 의하여 지원되었으며 지원해주신 각 당국에 감사의 뜻을 표합니다.

참고문헌

1. T.J. Mitchell and M. D. Morris, Baysesin Design and Analysis of Computer Experiments: Two Examples, *Statistica Sinica*, Vol.2,(1992), 359~379
2. Raymond H. Myers and Douglas C. Montgomery, *Response Surface Methodology –Process and Product Optimization Using Designed Experiments*, *John Wiley & Sons*, New York, USA,(1995)
3. Stella M. Clarke, H. Griebsch and Timothy W. Simpson, Analysis of Support Vector Regression For Approximation of Complex Engineering Analysis, *ASME Journal*, v.127 no.6, (2005), 1077-1087
4. M.J.D. Powell, Radial Basis Functions for Multivariable Interpolation: A review, *Algorithms for Approximation, Oxford University Press*, ,(1987) 143~167
5. Yongsik Shin, Yongbin Lee and Je-Seon Ryu, , Sequential Approximate Optimization Using Kriging Metamodels, *Transactions of the KSME*, Vol.29, No.9, ,(2005),1199~1208
6. Jin, R., Chen, W. and Sudjianto, A., On Sequential Sampling for Global Metamodeling in Enguneering Design, *Design Engineering Technical Conferences and Computeres and Information in Engineering Conference*, (2002).
7. Norman R.Draper and Harry Smith, Applied Regression Analysis ,*A Wiley-Interscience Publication*, New York, USA ,130~172
8. Jae Jun Jung ,Multiplicative Decomposition Method for Accurate Moment-Based Reliability Analysis, *Mechanical Design and Production Engineering*, *Hanyang Univesity* Seoul, Korea,(8)(2007) 26~36