한국지능시스템학회 논문지 2008, Vol. 18, No. 3, pp. 297-305

# 가우시안 프로세스 기반 함수근사와 서포트 벡터 학습을 이용한 레이더 및 강우계 관측 데이터의 융합

# Combining Radar and Rain Gauge Observations Utilizing Gaussian-Process-Based Regression and Support Vector Learning

유철상* · 박주영**

Chulsang Yoo and Jooyoung Park

* 고려대학교 건축사회환경공학과

** 고려대학교 제어계측공학과

## 요 약

최근들어, 커널 기법(kernel method)은 패턴 분류, 함수 근사 및 비정상 상태 탐지 등의 분야에서 상당한 관심을 끌고 있다. 특히, 서포트 벡터 머신(support vector machine)이나 커널 주성분 분석(kernel principal component analysis) 등의 방법론에서 커널의 역할은 매우 중요한데, 이는 고전적인 선형 머신이 비선형성을 효과적으로 다룰 수 있도록 일반화 해줄 수 있기 때문이다. 본 논문에서는 커널 기반 가우시안 프로세스(gaussian process) 함수근사 기법과 서포트 벡터 학습을 이용하여 레이더와 강우계의 관측 데이터를 융합하는 문제를 고려한다. 그리고, 국내의 강원, 경북 및 충북에 걸쳐있는 지역에 대한 레이더 자료 및 강우계 자료를 대상으로 하여 본 논문에서 고려하는 방법론들에 의해 데이터 융합을 수행한 결과를 제시하고, 성능비교를 수행한다.

**키워드** : 가우시안 프로세스, 서포트 벡터 학습, 커널 기법, 함수 근사, 데이터 융합, 수문학.

## Abstract

Recently, kernel methods have attracted great interests in the areas of pattern classification, function approximation, and anomaly detection. The role of the kernel is particularly important in the methods such as SVM(support vector machine) and KPCA(kernel principal component analysis), for it can generalize the conventional linear machines to be capable of efficiently handling nonlinearities. This paper considers the problem of combining radar and rain gauge observations utilizing the regression approach based on the kernel-based gaussian process and support vector learning. The data-assimilation results of the considered methods are reported for the radar and rain gauge observations collected over the region covering parts of Gangwon, Kyungbuk, and Chungbuk provinces of Korea, along with performance comparison.

**Key Words** : Gaussian process, Support vector learning, Kernel method, Function approximation, Data assimilation, Hydrology.

## 1. 서 론

최근 들어, 커널 기법(kernel methods)은 관련 이론이 정립되고 각종 응용 사례가 보고되면서, 지능시스템 분야에서 매우 주요한 도구 중 하나로 자리를 잡아가고 있다[1-12]. 특히, 서포트 벡터 머신(support vector machines)이나 커널 주성분 분석(kernel principal component analysis) 등의 방법론에서 커널(kernel)의 역할은 내우 중요한데, 이들은 기존의 선형 머신이 비선형성을 효과적으로 다룰 수 있도록 일반화해줄 수 있기 때문이다.

본 논문은 커널 기반 가우시안 프로세스(gaussian process) 함수 근사 기법과 서포트 벡터 학습 기법 각각에 대하여 조건부 병합 전략(conditional merging strategy)을 적용하여 수문학(hydrology) 분야에서 등장하는 레이더와 강우계의 관측 데이터를 융합하는 문제를 다룬다. 이를 위하여 본 논문에서는 최근에 관련 분야에서 소개되고 참고문헌[13]에서 기계 학습(machine learning) 관점에서 집대성된 바 있는 가우시안 프로세스 기반 함수 근사 기법과 서포트 벡터 학습을 이용한 함수 근사 기법을 커널 기법 관점에서 요약정리하고 이들 기법 각각을 주요 요소로 활용하여 국내의 특정 지역에서 관찰된 두 종류의 강우 관련 자료(즉, 레이더 자료와 강우계 자료)를 융합하는 문제에 적용한다. 본 논문에서 고려하는 응용문제는, 여러 종류의 계측기로부터 나오는 정보를 통합하여 강우량의 지역적 분포를 보다 정확

접수일자 : 2007년 7월 13일
완료일자 : 2007년 11월 19일
이 논문은 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2006-311-D00221)

하게 추정함을 시도한다는 측면에서 수문학 분야의 주요 관심 사항이 될 수 있다[14-16]. 가우시안 프로세스 기반 함수 근사 기법과 조건부 병합 전략을 활용한 강우 관련 데이터의 융합기법을 국내의 특정 지역을 대상으로 적용 및 기술한 후 서포트 벡터 머신 방법[1-3]과 조건부 병합 전략을 이용한 경우도 고려하고, 이들에 대한 성능비교도 수행한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 활용되는 주요 도구인 가우시안 프로세스 및 서포트 벡터 함수 근사 기법 등이 간략하게 소개된다. 3장은 본 논문의 핵심 주제를 다루는 부분으로써, 기계 학습(machine learning) 분야에서 최근에 주요한 도구로써 활용되고 있는 가우시안 프로세스를 이용한 함수 근사[13]의 관점에서, Ehrel에 의해 제시된 조건부 병합 전략[17]의 절차를 적용해보는 작업을, 국내의 강원, 경북 및 충북에 걸쳐있는 지역에 대해 확보된 레이더 자료 및 강우계 자료를 대상으로 수행한다. 그리고, 서포트 벡터 머신 방법이 조건부 병합 전략과 결합되어 적용되는 경우도 고려하고 성능비교도 수행해본다. 마지막으로, 4장에서는 결론과 향후 과제 등을 제시한다.

## 2. 기초 이론 소개

여러 종류의 계측기로부터 나오는 정보를 통합하여 강우량의 지역적 분포를 추정하는 작업은 수문학 관점에서 주요한 가치를 갖는 문제이다. 다음에서는, 본 논문에서 고려하는 강우 데이터 융합에 활용되는 주요 함수근사 도구인 가우시안 프로세스와 서포트 벡터 머신 기법을 간략하게 요약 정리 하도록 한다.

### 2.1 가우시안 프로세스를 이용한 함수근사

본 절에서는 본 논문에서 고려하는 주요 방법론 중 하나인 가우시안 프로세스 함수근사(Gaussian process regression) 기법의 내용을 기계 학습과 커널 기법 관점에서 간단히 정리하여 요약 설명하도록 한다. 이 기법에 관한 보다 자세한 설명을 위해서는 [13] 등의 참고문헌을 참조하면 된다. 가우시안 프로세스 $f(x)$는 $x \in R^d$를 인덱스로 갖는 랜덤변수 $f(x)$의 집합으로써, 이들 중 임의로 선택된 유한개의 랜덤변수 $f(x_1), \cdots, f(x_m)$가 반드시 결합 가우시안 밀도(joint Gaussian density)를 갖는 성질을 갖는 경우를 지칭한다[13]. 가우시안 프로세스는 평균함수(mean function)와 공분산 함수(covariance function)에 의해서 완전히 기술될 수 있으며, 이들은 각각 다음과 같이 정의된다:

$$m(x) = E[f(x)] \tag{1}$$

$$k(x,x') = E[(f(x)-m(x))(f(x')-m(x'))] \tag{2}$$

그리고, 가우시안 프로세스 $\{f(x)\}$가 평균함수 $m(x)$와 공분산함수 $k(x,x')$를 가짐은 다음과 같이 표기된다:

$$f(x) \sim GP(m(x), k(x,x')) \tag{3}$$

여기에서 소위 커널 트릭(kernel trick)을 이용하면, 특징 공간(feature space) 위에 정의되는 베이지안 선형 모델을 가우시안 프로세스로 취급할 수 있음을 쉽게 알 수 있다. 즉, $w$가 정규분포 $N(0, \Sigma_p)$를 선분포(prior distribution)로 갖고 $f(x)$가 $\phi(x)^T w$로 정의되는 확률 변수인 경우를 생각하자. 여기에서 $\phi(x)$는 특징 벡터로써, 입력 벡터 $x \in R^d$가 $\dim(F) \gg d$를 만족하는 고차원의 특징 공간 $F$로 매핑된 결과이다. 이때 $f(x)$의 기댓값은

$$E[f(x)] = E[\phi(x)^T w] = \phi(x)^T E[w] = 0 \tag{4}$$

를 만족하고, $f(x)$와 $f(x')$의 공분산은

$$E[f(x)f(x')] = E[\phi(x)^T w w^T \phi(x')] = \phi(x)^T E[ww^T]\phi(x') = \phi(x)^T \Sigma_p \phi(x') \tag{5}$$

를 만족한다. 따라서, 커널 트릭

$$k(x,x') = \phi(x)\, \Sigma_p \phi(x') \tag{6}$$

에 의해 커널 함수 $k$를 정의하면, 식 (5)에서 보여진 바와 같이 출력 공간 위에서의 공분산이 특징 공간 위에서 계산 과정이 생략된 채 $k(x,x')$에 의해 입력 공간 위에서 편리하게 직접 계산되는 모양새를 갖추게 된다. 이와 같이 정의되는 함수 $k$는 문맥에 따라 커널 혹은 공분산 함수로 다룰 수 있다. 가우시안 프로세스를 이용하여 함수 근사를 다룰 때에는 통상적으로 식 (1)의 평균함수 $m(x)$를 0으로 가정하는데, 이는 주어진 함수근사 문제에서 필요한 기저함수의 꼴이 미리 주어지지 않은 경우에는 합당한 선택이라고 볼 수 있다[13]. 학습 데이터가 지니고 있는 정보를 활용하여 시험 데이터 위에서의 참값을 근사하는 작업을 가우시안 프로세스 차원에서 수행하는 방법론은 다음과 같이 요약될 수 있다: 우선 $N$ 개의 입력 벡터로 구성되는 입력 벡터 집합 $X = \{x_n \in R^d\}^{N_n=1}$ 와 이에 대응하는 출력 값의 집합 $y = \{y_n \in R\}^{N_n=1}$, 그리고 이들 입출력 쌍을 동시에 표기하는 학습 데이터 집합 $D = \{(x_n, y_n)\}^{N_n=1}$를 고려하자. 평균함수 $m(x)=0$과 공분산함수 $k(x,x')$를 갖는 가우시안 프로세스 $f(x)$를 고려하면, 랜덤 벡터 $f = [f(x_1), \cdots, f(x_N)]^T$에 대한 확률밀도함수 정보는 다음과 같이 표기될 수 있다:

$$p(f \mid X) = N(f \mid 0, K(X)) \tag{7}$$

여기에서 $N(f \mid m, V)$는 평균이 $m$이고 공분산이 $V$인 다변량 정규분포를 의미하고, $K(X)$는 (i,j)번째 항을 위하여 공분산 함수 $k$에 입력 벡터 $x_i$와 $x_j$를 대입한 값 $k(x_i, x_j)$를 사용하는 $N \times N$ 행렬을 의미한다. 향후에는 간편한 표기를 위하여 $K(X)$는 $K$로 표기하기로 한다. 본 논문에서 선택한 공분산함수의 함수 꼴은 가우시안 프로세스를 이용한 함수근사 관련 연구에서 널리 사용되는 SE 형태(squared exponential form)로써, 다음과 같은 식으로 표현된다[13]:

$$k(x_i \cdot x_j) = \sigma_f^2 \times \exp[-\frac{1}{2}(x_i - x_j)^T diag(l)^{-2}(x_i - x_j)] \tag{8}$$

여기에서, $l = [l_1, \cdots, l_d]^T$은 입력 벡터 $x_i$등과 같은 차원을 갖고, 각 원소(entry)가 양수인 벡터이다. 그리고, 공분산 함수의 구체적인 함수 식을 결정해주는 파라메터 $\sigma_f$, $l$ 등은 하이퍼-파라메터 (hyper-parameters)로 지칭하고, 이들을 모두 모아 하이퍼-파라메터 벡터 $\theta$로 표기한다. 일반적으로 데이터 관측시 발생하는 오차를 위해서는 다음과 같은 가우시안 노이즈 모델을 고려한다:

$$p(y|f) = N(y|f, \sigma_n^2 I) \tag{9}$$

그리고, $p(f|X)$ 와 $p(y|f)$ 를 결합하면 다음과 같은 한계 우도(marginal likelihood)를 얻게된다:

$$p(y|X) = N(y|0, K+\sigma_n^2 I)$$

따라서, 주어진 학습 데이터에 대한 대수 한계 우도(log marginal likelihood)는 다음과 같이 전개될 수 있다[13]:

$$\begin{aligned}\log p(y|X) = &-\frac{1}{2} y^T (K+\sigma_n^2 I)^{-1} y \\ &-\frac{1}{2} log|K+\sigma_n^2 I| - \frac{N}{2}\log(2\pi)\end{aligned} \tag{10}$$

향후에 가우시안 프로세스 함수근사 GPR(gaussian process regression)의 적용과정에서 하이퍼-파라메터 $\theta$ 의 선택은 위의 대수 한계 우도 함수의 최적화를 통해 이루어지게 된다. 그리고, 시험용 입력 벡터 $x_*$ 에 대한 출력 값 $y_*$ 의 예측은 다변량 정규분포를 위한 조건부 확률 공식[13]을 적용하여 다음의 확률분포를 구한 후 이를 이용하여 구할 수 있다:

$$\begin{aligned}p(y_*|x_*, D) = N(&y_*|k_*^T(K+\sigma_n^2 I)^{-1} y, \\ &k_{**} - k_*^T(K+\sigma_n^2 I)^{-1} k_* + \sigma_n^2)\end{aligned} \tag{11}$$

위식에서 $k_*$ 과 $k_{**}$는 간편한 표기를 위하여 새롭게 도입된 변수로써, 각각 다음과 같이 공분산 함수 $k$를 사용하여 계산 한다:

$$k_* = [k(x_1, x_*), \cdots, k(x_n, x_*)]^T \tag{12}$$

$$k_{**} = k(x_*, x_*) \tag{13}$$

여기에서 주목해야할 점은, $y_*$를 위한 점 추정치로 사용될 평균 값 $k_*^T(K+\sigma_n^2 I)^{-1} y$은 $\alpha = (K+\sigma_n^2 I)^{-1} y$ 로 정의되는 계수 벡터(coefficient vector)

$$\alpha = [\alpha_1, \cdots, \alpha_N]^T \tag{14}$$

를 이용하여 정리하면

$$\hat{y}_* = \sum_{i=1}^{N} \alpha_i k(x_i, x_*) \tag{15}$$

꼴로 표기될 수 있다는 사실이다. 널리 알려진 바와 같이 식 (15)의 표현은 서포트 벡터 머신이나 커널 주성분 분석 등의 커널 기법에서 표현 정리(representer theorem)[3]의 결과로 등장하게 되는 비선형 함수이며, 커널의 선택여하에 따라 RBF 신경망(radial basis function neural networks) 이나 MLP(multi-layer perceptron) 등의 형태를 가질 수 있다[1-3].

### 2.2 서포트 벡터 학습을 이용한 함수 근사

본 논문에서는 커널에 기반을 둔 가우시안 프로세스와 조건부 병합 기법을 이용하여 레이더 및 강우계의 관측 데이터를 융합하는 방법에 대한 비교 대상으로, 최근에 지능 시스템 분야에서 각광을 받고 있는[1-3,10] 서포트 벡터 학습을 이용한 함수 근사 기법, 즉 SVR(support vector regression)을 고려한다. 현재 사용되고 있는 SVR 중 가장 널리 사용되는 기법으로는 입실론(epsilon) SVR 방법과 누(nu) SVR 방법이 있는데, 이들은 각각 줄여서 $\epsilon$-SVR 및 $\nu$-SVR로 표기된다. 다음에서는 본 논문의 교신저자가 작성한 관련 내용[10]에 따라 $\epsilon$-SVR과 $\nu$-SVR의 핵심을 간단히 요약소개하기로 한다. 우선 $\epsilon$-SVR의 목적은, 주어진 학습데이터 $\{(x_n, y_n)\}_{n=1}^N$에 대해 입실론-무반응 오차( $\varepsilon$-insensitive error)

$$|y_n - f(x_n)|_\epsilon \triangleq \max\{|y_n - f(x)| - \epsilon, 0\} \tag{16}$$

의 총합을 작게 하는 매끄러운 근사함수인

$$f(x) = < w, \phi(x) > + b \tag{17}$$

를 찾는 것으로 표현될 수 있다. 여기에서, $\phi(x)$는 앞 절에서와 마찬가지로 입력 벡터 x를 고차원 특징공간 $F$로 매핑한 결과이다. 그리고, 위에서 언급한 $\epsilon$-SVR의 목적은, 다음과 같은 최적화 문제를 풀어서 달성할 수 있다[1-3]:

$$\begin{aligned}\min\ &\frac{1}{2}|w|^2 + C\sum_{n=1}^{N}(\xi_n + \xi_n^*) \\ s.t.\quad &y_n - (< w, \phi(x_n) > + b) \le \epsilon + \xi_n \\ &(< w, \phi(x_n) > + b) - y_n \le \epsilon + \xi_n^* \\ &\xi_n, \xi_n^* \ge 0, \ \forall n\end{aligned} \tag{18}$$

여기에서, $C > 0$는 함수 $f$ 의 완만함(flatness)과 학습 데이터에 대한 오차 $\xi_n, \xi_n^*$ 사이의 상대적 중요도를 결정하는 상수이다. 이러한 최적화 문제를 라그랑제 쌍대문제(Lagrange dual problem)을 활용하여 풀어주면 최종적으로 $w = \sum_{n=1}^{N}(\alpha_n - \alpha_n^*)\phi(x_n)$ 가 얻어지고, 이를 (17)에 대입한 후 커널 트릭을 적용하면 다음과 같은 근사 함수 $f$가 얻어진다:

$$f(x) = \sum_{n=1}^{N}(\alpha_n - \alpha_n^*)k(x_n, x) + b \tag{19}$$

단, 이 식에서 커널 함수의 계수를 구성하는 $\alpha_n, \alpha_n^*$는 다음의 쌍대 최적화 문제로부터 얻어지는 해이다[1-3].

$$\begin{aligned}\max\ &-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(\alpha_n - \alpha_n^*)(\alpha_m - \alpha_m^*)k(x_n, x_m) \\ &+\sum_{n=1}^{N}(\alpha_n - \alpha_n^*)y_n - \sum_{n=1}^{N}(\alpha_n + \alpha_n^*)\epsilon \\ s.t.\quad &\sum_{n=1}^{N}(\alpha_n - \alpha_n^*) = 0 \\ &\alpha_n, \alpha_n^* \in [0, C], \ \forall n\end{aligned} \tag{20}$$

한편, 본 논문에서 비교 대상으로 활용할 $\nu$-SVR은 누(nu) 값의 선택에 따라 식 (16)에 등장하는 $\epsilon$ 튜브의 크기가 자동으로 정해지는 방향으로 위의 $\epsilon$-SVR을 확장한 방법이다. 다음에서는 [10]의 내용에 따라 $\nu$-SVR의 주요 사항을 간략하게 기술하도록 한다. 이에 관한 보다 저 상세한 기술을 위해서는 [1-3]을 참고하면 된다. 널리 알려진 바와 같이, $\epsilon$-SVR 학습방법이 갖는 문제점 중 하나는, 학습이 수행되기 전에 $\epsilon$의 값을 미리 정해주어야 한다는 것이다. 그러나, 대부분의 함수 근사 문제에서는 주어진 학습 데이터에 합당한 무반응 오차 범위 $\epsilon$값을 미리 알 수 없다. 이러한 문제를 해결하기 위해서 $\nu$-SVR 학습방법에서는 $\epsilon$의 값을 최적화 문제의 변수들 중 하나로 인식하고, 다른 변수

들과 함께 가능한 한 작은 $\epsilon$의 값을 갖도록 최적화 문제에 포함시켜 표현함으로써, 함수 근사를 위한 최적화 문제를 해결함과 동시에 최적의 $\epsilon$값을 구하는 방법을 취한다. 또한, $\nu$-SVR 학습방법은 v값의 변화에 따라 서포트 벡터의 개수를 조절할 수 있는 장점이 있다. $\nu$-SVR 기법을 위한 최적화 문제는 주어진 상수 $\nu \in (0,1]$ 에 대해서, 다음과 같이 구성된다:

$$\begin{aligned} \min \quad & \frac{1}{2}|w|^2 + C\sum_{n=1}^{N}(\xi_n + \xi_n^* + \nu\epsilon) \\ s.t. \quad & y_n - (<w,\phi(x_n)>+b) \le \epsilon + \xi_n \\ & (<w,\phi(x_n)>+b) - y_n \le \epsilon + \xi_n^* \\ & \xi_n, \xi_n^*, \epsilon \ge 0\,,\ \forall n \end{aligned} \tag{21}$$

이 문제의 목적 함수와 제약 조건으로부터, 문제 (21)에 대응하는 라그랑제 함수는 다음과 같음을 알 수 있다:

$$\begin{aligned} L = & \frac{1}{2}|w|^2 + C\sum_{n=1}^{N}(\xi_n + \xi_n^* + \nu\epsilon) - \gamma\epsilon \\ & - \sum_{n=1}^{N}\alpha_n(\epsilon + \xi_n - y_n + <w,\phi(x_n)>+b) \\ & - \sum_{n=1}^{N}\alpha_n^*(\epsilon + \xi_n^* + y_n - <w,\phi(x_n)>-b) \\ & - \sum_{n=1}^{N}(\eta_n\,\xi_n + \eta_n^*\,\xi_n^*) \end{aligned} \tag{22}$$

여기에서, 각 라그랑제 승수는 조건

$$\alpha_n, \alpha_n^*, \eta_n, \eta_n^*, \gamma \ge 0 \tag{23}$$

을 만족하고, 최적 해에 대한 안장점 조건을 이용하여 라그랑제 함수의 주 변수(primal variable)를 소거하면 다음과 같은 쌍대 문제를 얻게 된다.

$$\begin{aligned} \max \quad & -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(\alpha_n - \alpha_n^*)(\alpha_m - \alpha_m^*)k(x_n,x_m) \\ & + \sum_{n=1}^{N}(\alpha_n - \alpha_n^*)y_n \\ s.t. \quad & \sum_{n=1}^{N}(\alpha_n + \alpha_n^*) \le CN\nu,\ \sum_{n=1}^{N}(\alpha_n - \alpha_n^*) = 0 \\ & \alpha_n, \alpha_n^* \in [0, C]\,,\ \forall n \end{aligned} \tag{24}$$

쌍대 문제 (23)를 풀면, 최적의 $\alpha_n$와 $\alpha_n^*$가 구해지고, 이에 따르는 최적의 $b$와 $\epsilon$ 값은 Kuhn-Tucker 조건을 이용하여 얻을 수 있다. 그리고, $\nu$-SVR 학습방법을 통하여 얻은 근사 함수 $f$ 는, $\epsilon$-SVR 기법의 경우와 마찬가지로 다음과 같이 표현할 수 있게 된다:

$$\begin{aligned} f(x) &= <w,\phi(x)>+b \\ &= \sum_{n=1}^{N}(\alpha_n - \alpha_n^*)\,k(x_n,x) + b \end{aligned} \tag{25}$$

## 3. 가우시안 프로세스, SVM 및 조건부 병합 기법을 활용한 강우 데이터의 융합

여러 종류의 계측기로부터 나오는 정보를 통합하여 강우량의 지역적 분포를 추정하는 작업은 수문학 분야의 중요한 연구 주제 중 하나이다. 본 논문에서는 레이더를 이용한 강우 필드(rainfall field)의 관측 값과 강우계 관측 값을 효과적으로 통합하는 데이터 융합 문제를 고려한다. 널리 알려진 바와 같이, 레이더를 이용한 강우 필드의 관측은 여러 가지 종류의 오류를 갖지만[18-19], 실제 강우 필드의 공분산 구조(covariance structure)를 어느 정도 표현할 수 있는 장점을 보유하게 된다. 따라서, 레이더를 이용한 관측 정보는 미지의 바이어스 항에 의한 영향 등으로 말미암아 그 자체로는 부정확한 한계를 갖지만, 강우계 측정치를 단순히 내삽(interpolation)하는 방법들이 가질 수밖에 없는 공간적 정보의 결핍을 보완하는 데 유용하게 사용될 수 있다. 뿐만 아니라, 여기에서 한 단계 더 나아가서 강우계 위치 근처에서는 강우계 관측 값을 존중하면서 전체적으로는 올바른 공간적 구조를 갖도록 하는 방향으로 활용할 경우, 효과적인 강우 필드의 추정을 가능하게 할 수 있다. 이러한 취지에서 Ehrel에 의해 연구된 조건부 병합 전략[17]에서는, 관측된 데이터의 융합에 필요한 정보를 추출하기 위하여 통상적인 크리깅(ordinary kriging, 이에 관한 상세한 사항을 위해서는 [20] 등을 참조하기 바람)을 이용하는 절차를 소개한 바 있다. 본 논문에서는, 최근에 [13]을 통하여 집대성된 바 있는 가우시안 프로세스의 기계 학습적 관점을 Ehrel의 조건부 병합 전략[17]과 접목하여, 국내의 강원, 경북 및 충북에 걸쳐있는 지역에 관한 레이더 자료 및 강우계 자료(그림 1 참고)를 대상으로 하여 적용해보고, 이와 함께 서포트 벡터 머신에 의한 함수 근사와 조건부 병합 전략을 함께 사용한 경우와의 성능 비교도 수행해보도록 한다.

본 논문에서 고려하는 강우 자료가 취득된 곳의 위치는 그림 1에서 보여진 바와 같이 강원, 경북, 충북에 걸쳐있는 지역이다. 이 그림에서, 레이더 강우 자료가 얻어지는 부분은 격자 형태로 표기되어 있고, 건설교통부(MOCT)에 의해 지상강우량이 측정되는 강우계 위치는 'o' 모양으로 표기되어 있다. 본 논문에서 고려하는 구체적인 강우사상은 2003년 6월 27일 13시(LST)에 관찰된 대류성 강우 사상으로써, 각 위치에서 측정된 시간당 강우량은 그림 2(강우계 자료) 및 그림 3(레이더 자료)의 삼차원 그래프와 같다. 이들 그래프에서 보여진 강우량의 단위는 [mm/hr]이며, 본 논문에서 활용한 레이더 관련 데이터는 그림 3에서 보여진 바와 같이, 그림 1의 레이더 격자들 중 일정한 간격으로 샘플링 된 전체의 $1/5^2$ 부분이다.

2장에서 소개된 가우시안 프로세스 기반 함수근사 기법을 그림 2의 강우계 자료에 적용하면 레이더 격자가 위치하는 각 그리드에 대해 일정한 강우 추정치를 얻을 수 있다. 추정치를 얻는 과정에서 하이퍼-파라메터 $\sigma_f, l_1, l_2$ 및 노이즈 표준편차 $\sigma_n$ 은, 식 (15)의 대수 한계 우도 함수에 비선형 최적화 기법([13] 참조)을 적용하여 얻을 수 있는데, MOCT 강우계 자료를 학습 데이터로 사용하여 최적화를 수행한 결과 다음과 같은 결과를 얻었다:

$$\begin{aligned} &[\log\sigma_f, \log l_1, \log l_2, \log\sigma_n] \\ &= [1.8764\,,\ 2.2292\,,\ 2.0431\,,\ -4.0935] \end{aligned} \tag{26}$$

그림 4에는 위에서 언급한 강우 추정치가 강우계 위치에서의 실측치('o' 표기)와 함께 보여 졌으며, 이 추정치는 참값과의 괴리로 표현되는 오차 $\epsilon_G(s)$ 를 수반하게 된다. 따라서, 이와 같은 상황은 다음과 같은 식으로 요약될 수 있다[21]:

$$T(s) = \hat{G}(s) + \epsilon_G(s) \tag{27}$$

여기에서, $T(s)$는 그리드 위치 $s$에서의 참 강우량(true rainfall field)을 의미하고, $\hat{G}(s)$는 가우시안 프로세스 기반 함수근사 기법을 강우계 실측치에 적용하여 구한 그리드 $s$에서의 추정 강우량(즉, $T(s)$의 추정값)이다. 그리고, $\epsilon_G(s)$는 위치 $s$에서의 참값과 추정값 사이의 오차를 나타낸다. 다음 단계에서는 가우시안 프로세스 기반 함수근사 기법을 각 강우계 위치에서의 (추정된) 레이더 자료에 적용하여, 레이더 격자가 위치하는 각 그리드에 대해 일정한 레이더 강우 추정치를 구한다. 그림 5에는 이렇게 구한 레이더 강우 추정치가 강우계 위치에서의 레이더 자료('+' 표기)와 함께 보여졌으며, 이러한 추정치는 오차 $\epsilon_R(s)$를 수반하게 된다. 그리고, 이러한 상황은 식 (27)의 경우와 마찬가지로 다음과 같이 표현될 수 있다[21]:

$$R(s) = \hat{R}(s) + \epsilon_R(s) \tag{28}$$

여기에서, $R(s)$는 그리드 위치 $s$에서의 참 레이더 강우값을 나타내고 $\hat{R}(s)$는 가우시안 프로세스 기반 함수근사 기법을 강우계 위치의 레이더 자료에 적용하여 구한 그리드 $s$에서의 레이더 강우 추정량(즉, $R(s)$의 추정값)이다. 그리고, $\epsilon_R(s)$는 참값 $R(s)$와 추정값 $\hat{R}(s)$ 사이의 오차를 나타내는데, 이 경우 $R(s)$의 참값이 실측된 레이더 강우자료로부터 직접 구해질 수 있는 관계로 식 (28)의 $R(s)$, $\hat{R}(s)$ 및 $\epsilon_R(s)$는 모두 직접 구할 수 있는 항들이 된다. 반면에, 식 (27)의 경우에는 $\hat{G}(s)$만 직접 구할 수 있는 항이고, $T(s)$와 $\epsilon_G(s)$는 미지의 항으로 남게 된다. 조건부 병합 전략의 핵심[17,21]은 (27)의 참값 $T(s)$를 추정하기 위하여

$$M(s) = \hat{G}(s) + \epsilon_R(s) \tag{29}$$

를 사용할 수 있다는 것인데, 이 경우 식 (29)의 좌변은 병합(merging) 과정을 통하여 구한 $T(s)$에 대한 추정치를 나타내는 랜덤 변수명을 의미하고, 우변의 두 항은 각각 식 (27)과 (28)의 요소로써 위에서 언급된 바와 같이 직접 구할 수 있는 항들이다. 따라서, 식 (29)을 통해 계산되는 $M(s)$가 바로 가우시안 프로세스와 조건부 병합 전략이 결합되어 제공하는 추정치가 되는 것이다. 이러한 상황은 오차에 관한 정보를 표현하는 중간 단계를 거치면 보다 용이하게 설명될 수 있다. 예를 들어 그림 6에서는 식 (28)에서 오차 항 $\epsilon_R(s)$가 구해지는 과정이 보여졌는데, 여기에서 각 점 $R(s)$과 메쉬 그리드 부분$\hat{R}(s)$을 연결하는 선분 부분이 오차의 크기를 나타내고 있다. 식 (29)에서 최종적인 추정치를 구하기 위해 $\hat{G}(s)$에 $\epsilon_R(s)$ (즉, 선분 크기 부분)을 더해주는 과정은 그림 7에 보여졌다. 가우시안 프로세스와 조건부 병합을 활용하여 최종적으로 구한 추정치는 그림 8에 강우계 실측치(즉, 강우계 위치에서의 참 강우량 값)과 함께 보여졌다. 그림에 잘 나타나 있듯이, 최종 결과에는 각 실측치에 대한 매끄러운 내삽과 레이더 자료가 제공하는 공간적 변화 정보가 모두 고려되어 있다. 마지막으로, 해당 지역의 각 위치가 갖는 강우량에 대한 등고선 정보는 그림 9에 보여졌다.

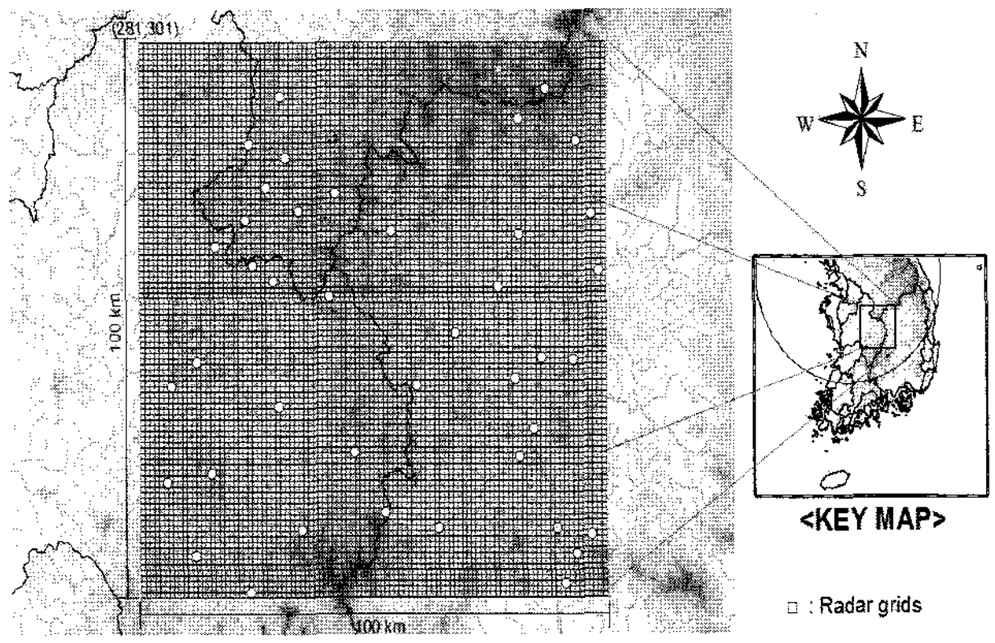


그림 1. 레이더 자료의 공간분포 및 MOCT 강우계 위치.
Fig. 1. Extent of radar data and locations of MOCT rain gauges.

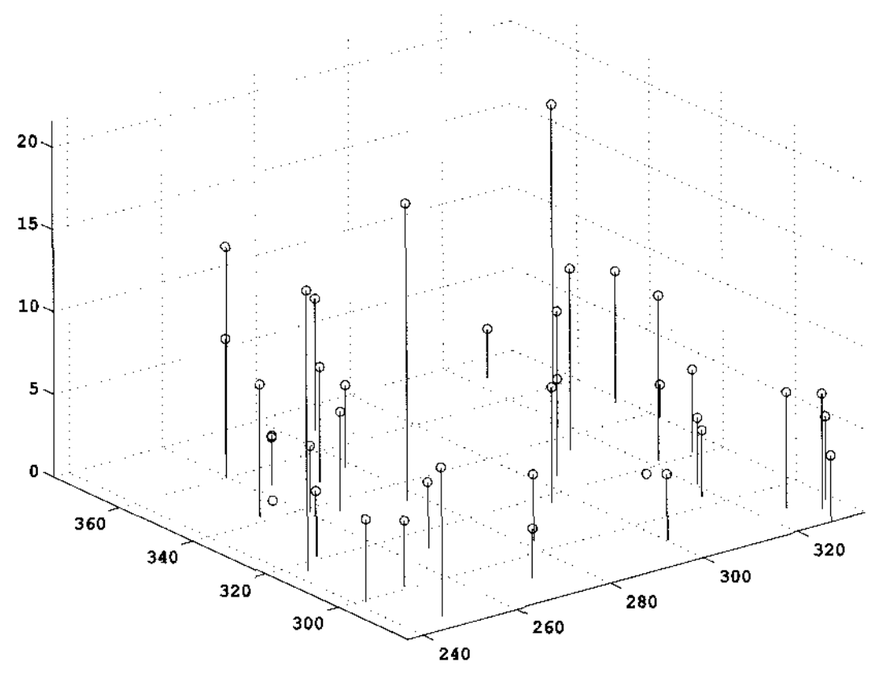


그림 2. 건설교통부(MOCT) 강우계에 의해 관측된 강우량 자료.
Fig. 2. The rainfall data observed by MOCT rain gauges.

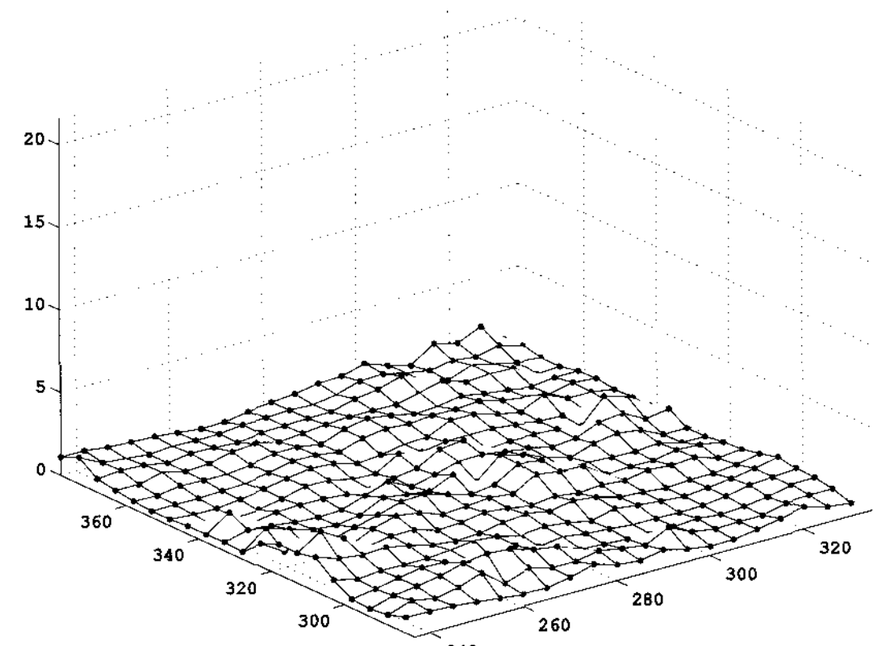


그림 3. 일정 간격 그리드에 위치한 레이더에 의해 관측된 강우량 자료.
Fig. 3. The rainfall data observed by radar on a regular grid.

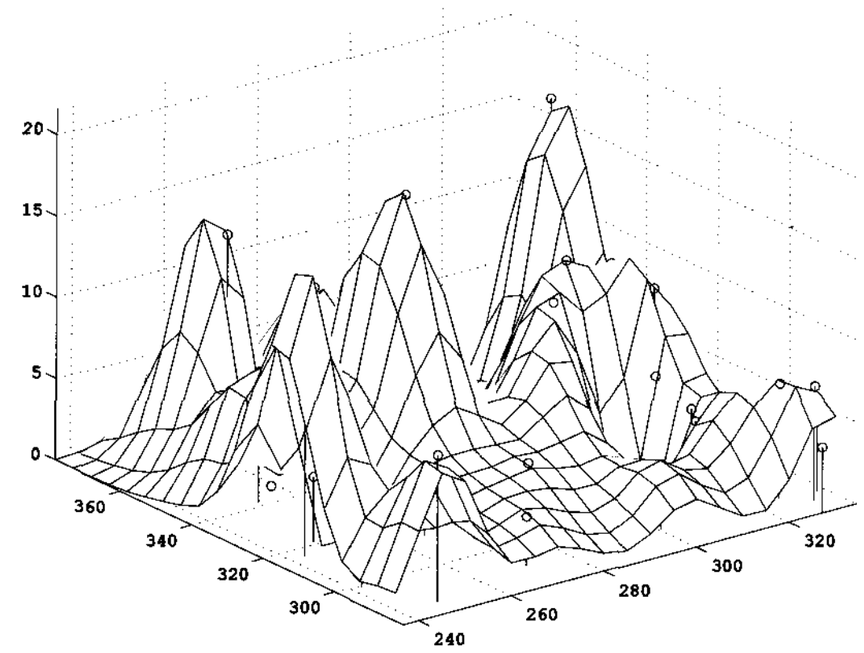


그림 4. 가우시안 프로세스 기반 함수근사를 이용하여 구한 레이더 그리드 위치에서의 강우량 추정값.

Fig. 4. Rainfall estimate on the radar grid obtained by the gaussian-process-based regression.

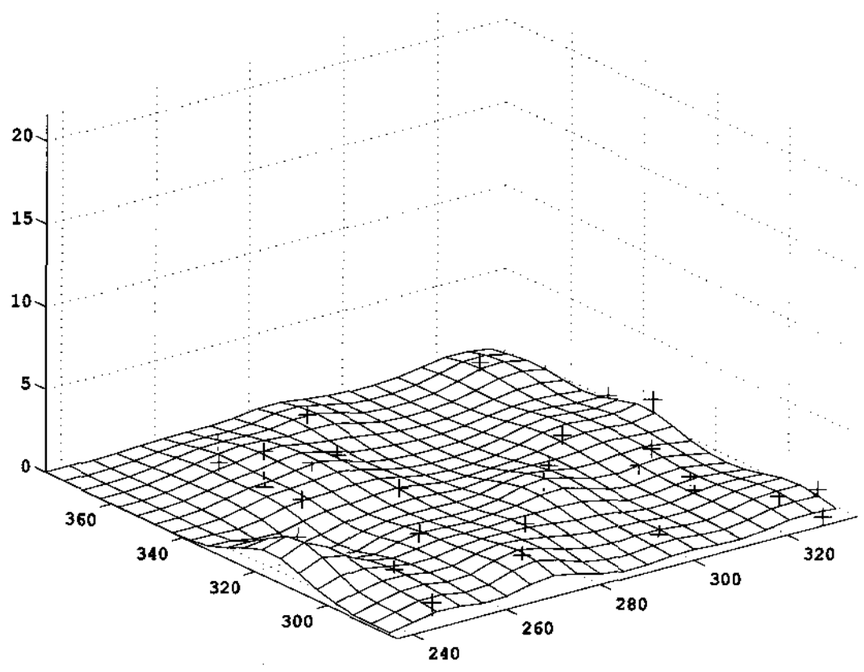


그림 5. 가우시안 프로세스 기반 함수근사를 이용하여 구한 레이더 그리드 위에서의 레이더 강우 추정값.

Fig. 5. Radar rainfall estimate on the radar grid obtained by the gaussian-process-based regression.

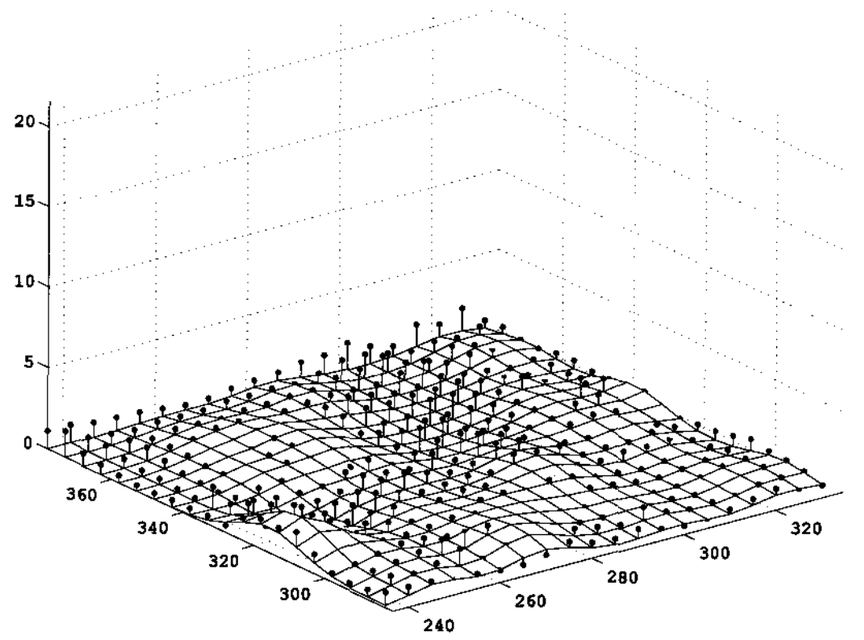


그림 6. 식 (28)로부터 오차항 $\epsilon_R(s)$ 를 추출해내는 과정.

Fig. 6. The process for obtaining the error term $\epsilon_R(s)$ from equation (28).

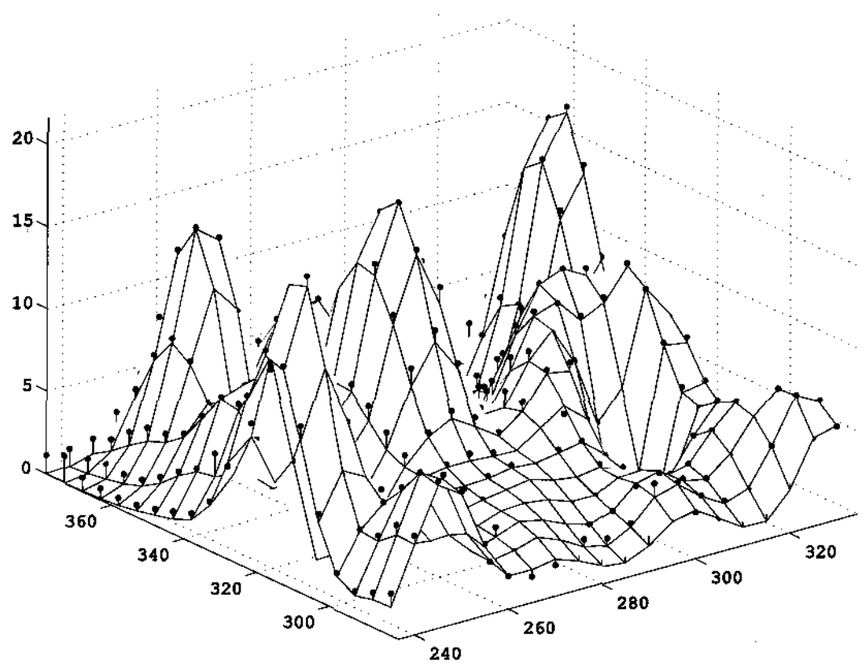


그림 7. $\hat{G}(s) + \epsilon_R(s)$ 를 계산하는 과정.

Fig. 7. The process for computing $\hat{G}(s) + \epsilon_R(s)$.

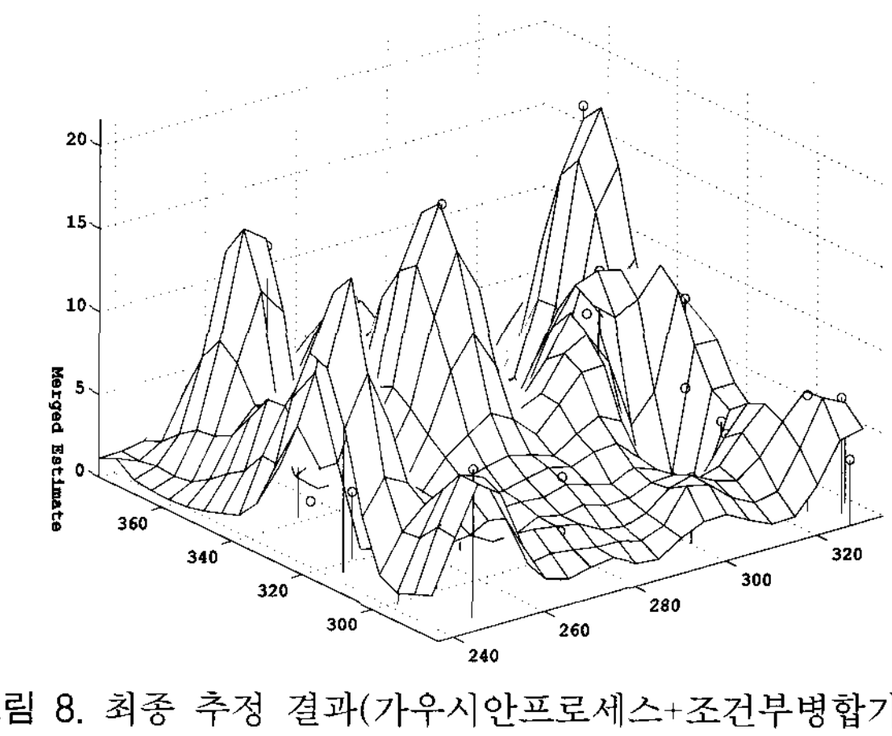


그림 8. 최종 추정 결과(가우시안프로세스+조건부병합기법) 및 강우계 실측치.

Fig. 8. Final estimation results (from Gaussina process + conditional merging strategy) along with rainfall observations.

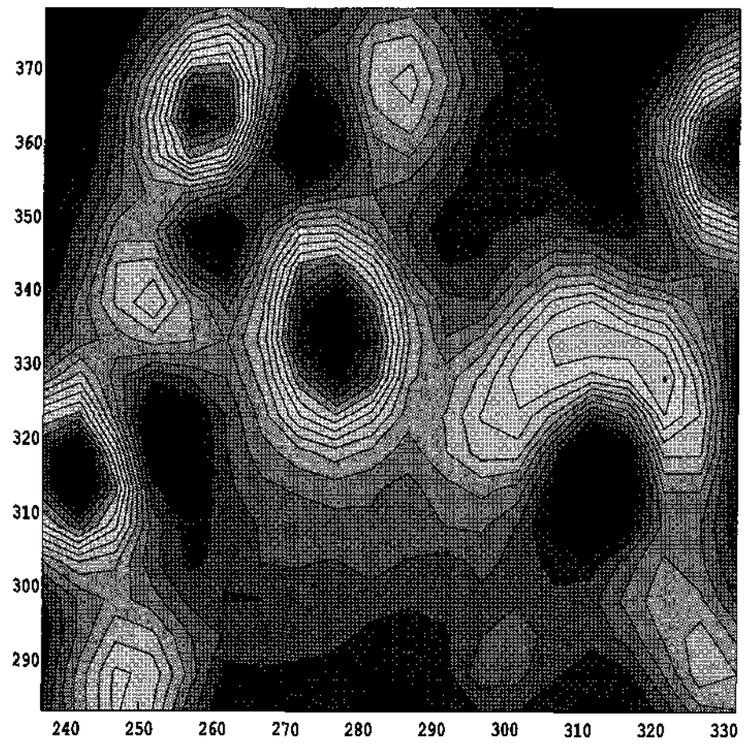


그림 9. 추정된 강우 필드에 대한 등고선 정보.

Fig. 9. Contour contents for the estimated rainfall field.

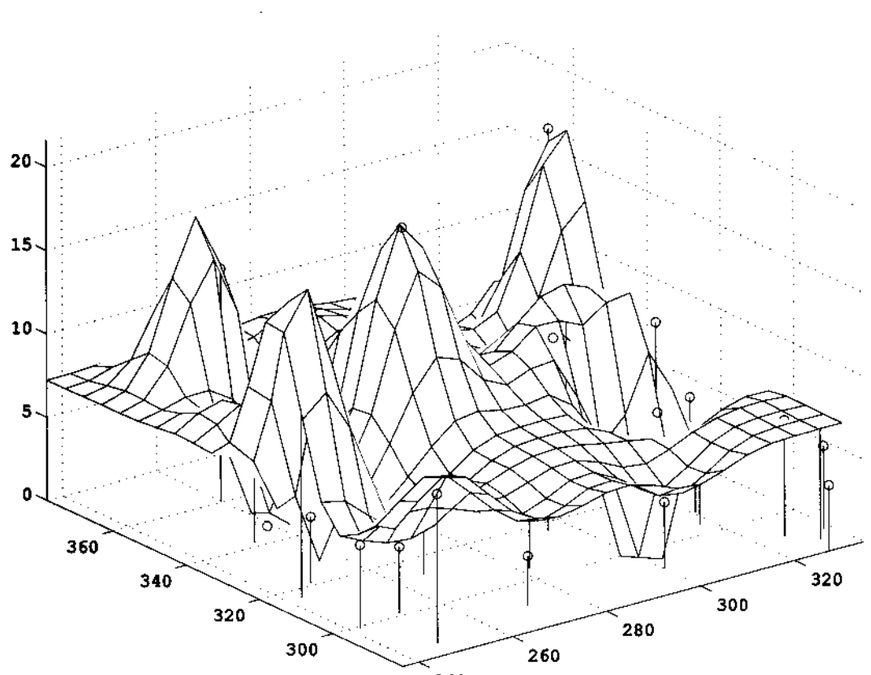


그림 10. SVM을 이용하여 구한 레이더 그리드 위에서의 강우량 추정.

Fig. 10. Rainfall estimate on the radar grid obtained by the SVM.

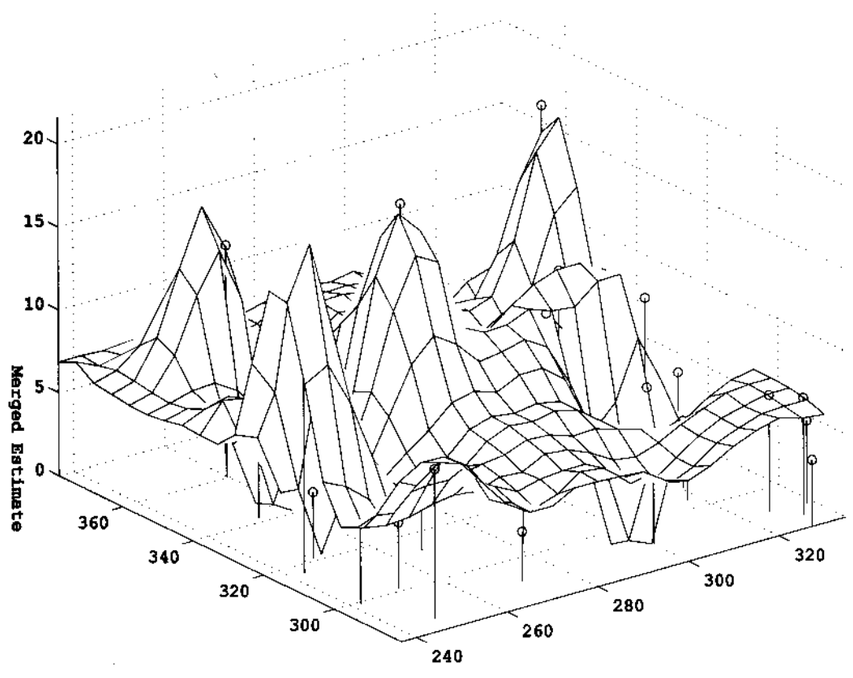


그림 11. 최종 추정 결과(SVM+조건부병합기법) 및 강우계 실측치.

Fig. 11. Final estimation results (from SVM + conditional merging strategy) along with rainfall observations.

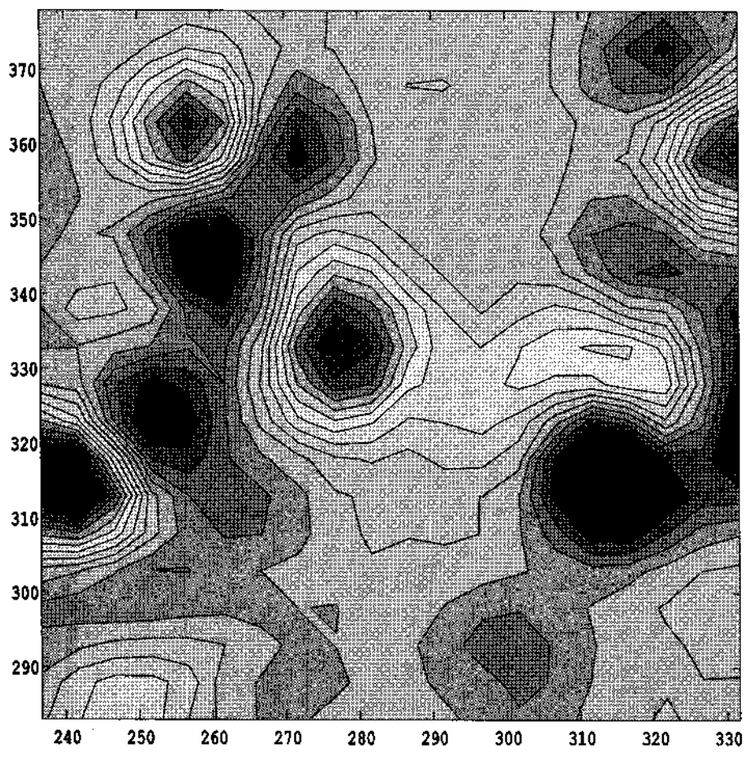


그림 12. 추정된 강우 필드에 대한 등고선 정보.

Fig. 12. Contour contents for the estimated rainfall field.

이상에서 가우시안 프로세스와 조건부 병합을 활용한 강우계 및 레이더 관측 데이터의 융합기법을 국내의 특정 지역에서 관측된 자료에 적용시켜 보았다. 다음에서는 성능 비교를 위하여 가우시안 함수 $k(x,z)=\exp(-|x-z|^2/2\sigma^2)$ 를 커널로 사용하는 서포트 벡터 학습[1-3] 기반 함수근사 결과를 고려한다. 다음의 비교 과정에서는, 서포트 벡터 학습[1-3]에 조건부 병합 전략을 적용하는 방법을 고려할 예정이다. 참고로, 조건부 병합 전략을 거치지 않고, 강우계의 자료와 레이더의 자료를 동시에 학습 데이터로 고려하는 경우는 해석 상 다소 불합리한 요소를 안고 있다고 할 수 있다. 그 이유는 강우계 실측치는 소수이지만 매우 정교한 관측 결과를 제공할 수 있는 반면에 레이더 자료는 다수인 반면에 매우 큰 바이어스 항을 가지므로, 이 두 종류의 자료를 동급의 학습 데이터로 취급하여 기존의 SVM 기법을 그대로 이 경우에 적용될 경우 학습 데이터 중 다수를 차지하는 부정확한 레이더 자료를 중심으로 한 함수근사 결과가 나오는 것은 피할 수 없는 이치이다. 따라서, 이러한 종류의 데이터 융합 문제에 조건부 병합 전략과 같은 부분은 상당히 큰 역할을 할 수밖에 없음은 자명하다고 할 수 있다. 본 논문에서는 가우시안 프로세스에 조건부 병합 전략을 결합한 것과 같은 방식으로 서포트 벡터 학습과 조건부 병합 전략을 결합하는 과정을 거친 후 그림1-3의 데이터의 융합에 적용하여 보았다. 이 과정에서는, 식 (27)-(29)에 대응하는 다음 식 (30)-(32)이 도입된 후 이들이 서포트 벡터 학습과 조건부 병합 전략의 결합과정에 활용되었다:

$$T(s)=\hat{f}(s)+\epsilon_f(s) \tag{30}$$

여기에서, $T(s)$ 는 그리드 위치 $s$ 에서의 참 강우량(true rainfall field)을 의미하고, $\hat{f}(s)$ 는 $\nu$-SVR 기법을 강우계 실측치에 적용하여 구한 그리드 $s$ 에서의 추정 강우량(즉, $T(s)$ 의 추정값)이다(이 단계에서 사용된 서포트 벡터 학습의 파라메터는 다음과 같음: $C=50$ , $\nu=0.5$, $\sigma^2=50$). 그리고, $\epsilon_f(s)$ 는 위치 $s$ 에서의 참값과 추정값 사이의 오차를 나타낸다. 다음 단계에서는 서포트 벡터 학습 기법을 각 강우계 위치에서의 (추정된) 레이더 자료에 적용하여, 레이더 격자가 위치하는 각 그리드에 대해 일정한 레이더 강우 추정치를 구한다(이 단계에서 사용된 서포트 벡터 학습의 파라메터는 다음과 같음: $C=50$ , $\nu=0.5$, $\sigma^2=0.5$). 이러한 추정치는 오차 $\epsilon_R(s)$ 를 수반하게 된다. 그리고, 이러한 상황은 식 (27)의 경우와 마찬가지로 다음과 같이 표현될 수 있다:

$$g(s)=\hat{g}(s)+\epsilon_g(s) \tag{31}$$

여기에서, $g(s)$ 는 그리드 위치 $s$ 에서의 참 레이더 강우값을 나타내고 $\hat{g}(s)$ 는 가우시안 프로세스 기반 함수근사 기법을 강우계 위치의 레이더 자료에 적용하여 구한 그리드 $s$ 에서의 레이더 강우 추정량(즉, $g(s)$ 의 추정값)이다. 그리고, $\epsilon_g(s)$ 는 참값 $g(s)$ 와 추정값 $\hat{g}(s)$ 사이의 오차를 나타내는데, 이 경우 $g(s)$ 의 참값이 실측된 레이더 강우자료로부터 직접 구해질 수 있는 관계로 식 (30)의 $g(s)$ , $\hat{g}(s)$ 및 $\epsilon_g(s)$ 는 모두 직접 구할 수 있는 항들이 된다. 따라서, 조건부 병합 전략[17,21]을 따라서(30)의 참값 $T(s)$ 는 다음 식을 통하여 추정될 수 있다:

$$M_{svr}(s) = \hat{g}(s) + \epsilon_g(s) \tag{32}$$

이상에서 설명된 각 과정의 주요 결과들은 그림 10-12에서 보여졌는데, 이들은 그림 4, 8-9에 대응하는 절차를 나타낸다. 마지막으로 별도의 시험 데이터 집합을 대상으로 하여 함수 근사 결과로 구해진 추정치 $\hat{y}$ 의 정확도를 비교한 결과를 표 1에 요약하였다. 이 결과는 같은 시간 같은 장소에서 취득된 AWS 강우계 실측치[23]를 시험 데이터로 하고 평균자승오차(mean squared error)

$$MSE = \frac{1}{N_{test}} = \sum_{n=1}^{N_{test}} (y_n - \hat{y}_n)^2 \tag{33}$$

와 평균절대오차(mean absolute error)

$$MAE = \frac{1}{N_{test}} = \sum_{n=1}^{N_{test}} |y_n - \hat{y}_n|$$

를 성능지수로 하여 얻어졌는데, 표 1에서 보여진 바와 같이, 가우시안 프로세스와 조건부 병합 전략을 채용한 경우와 서포트 벡터 학습과 조건부 병합 전략을 채용한 경우 모두 조건부 병합 전략없이 얻어진 결과에 비해 우수한 추정결과를 보여줌을 알 수 있었다. 그리고, 가우시안 프로세스와 서포트 벡터 학습 사이의 성능 비교에서는 조건부 병합 전략의 채용 여부에 관계없이 서포트 벡터 학습이 보다 더 좋은 결과를 제공함을 관찰할 수 있었다. 본 논문에서는, 강우관련 자료융합분야에서 가우시안 프로세스 기법의 일종으로 볼 수 있는 크리깅 기법[20]이 더러 사용되어 왔음을 고려하여, [13]의 가우시안 프로세스 기반 함수 근사기법 관련 부분에 관해 SVM에 못지않은 분량으로 비교적 깊이 있게 기술하였다. 보다 폭넓은 자료를 바탕으로 한 각 방법론의 성능비교는 의미 있는 향후 연구가 될 것이다.

## 4. 결론 및 토의

본 논문에서는 최근에 기계학습 분야에서 주목을 끌고 있는 커널 기반 방법론에 속하는 가우시안 프로세스와 서포트 벡터 머신의 각각에 조건부 병합 전략을 결합하여, 레이더를 이용한 강우 필드의 관측 값과 강우계 관측 값을 효과적으로 통합하는 문제를 고려하였다. 강원, 경북, 충북을 포함하는 국내의 특정 지역에서 수집된 강우 자료에 대해 위의 방법론 각각을 적용한 결과 그림 4-9 및 그림 10-12와 같은 결과를 보여주었다. 그리고 이들 그림을 통해 관찰된 내용과 표 1의 성능 비교로부터, 본 논문에서 고려하는 융합적 방법론들은 조건부 병합 전략이 사용되지 않은 기존의 함수 근사 기법에 비해 우수한 강우 데이터 융합 결과를 제공함을 알 수 있었다. 본 논문에서 고려한 내용과 관련하여 추가로 수행될 필요가 있는 연구 주제로는, 보다 폭넓은 자료를 바탕으로 한 광범위한 성능 비교와 조건부 병합 전략을 보다 서포트 벡터 머신에 어울리는 형태로 변형시키는 문제 등을 들 수 있다.

표 1. AWS 강우계 실측치를 시험 데이터로 사용한 성능 비교.

Table 1. Performance comparison utilizing the AWS rainfall gauge observations as test data.

| 방법론 | MSE | MAE |
|---|---|---|
| 가우시안 프로세스와 조건부 병합 전략을 사용한 경우 | 35.87 | 4.701 |
| 가우시안 프로세스만 사용한 경우 | 38.38 | 4.890 |
| 서포트 벡터 학습과 조건부 병합 전략을 사용한 경우 | 14.70 | 3.116 |
| 서포트 벡터 학습만 사용한 경우 | 26.51 | 3.648 |

## 참 고 문 헌


[1] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[2] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.

[3] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT Press, 2002.

[4] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181-201, 2001.

[5] J. T. Kwok, "The evidence framework applied to support vector machines," *IEEE Transactions on Neural Networks*, vol. 11, no. 5, pp. 1162-1173, 2000.

[6] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, and A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, pp. 1443-1471, 2001.

[7] B. Schölkopf, A. Smola, R. Williamson, and P. L. Bartlett. "New support vector algorithms," *Neural Computation*, vol. 12, no. 5, pp. 1207-1245, 2000.

[8] 이한성, 임영희, 박주영, 박대희, "SVM과 클러스터링 기반 적응형 침입탐지 시스템", *한국 퍼지 및 지능시스템 학회 논문지*, 13권 2호, pp. 237-242, 2003년 4월.

[9] 박주영, 임채환, "비정상 상태 탐지를 위한 서포트 벡터 학습", *한국 퍼지 및 지능시스템 학회 논문지*, 13권 3호, pp. 266-274, 2003년 6월.

[10] 김영일, 조원희, 박주영, "정해진 기저함수가 포함되는 Nu-SVR 학습방법", *한국 퍼지 및 지능시스템 학회 논문지*, 13권 3호, pp. 316-321, 2003년 6월.

[11] J. Park, J. Kim, H. Lee, D. Park, "One-class support vector learning and linear matrix inequalities," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 3, no. 1, pp.

100-104, June 2003.
[12] J. Park and D. Kang, "A Modified approach to density-induced support vector data description," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 7, no. 1, pp. 1-6, March 2007.
[13] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.
[14] G. De Marsily, *Quantitative Hydrogeology*, Academic Press, 1986.
[15] D. J. Seo, and J. A. Smith, "Rainfall estimation using raingages and radar. A Bayesian approach: 1. Derivation of estimators," *Stochastic Hydrology and Hydraulics*, vol. 5, pp. 17 - 29, 1991.
[16] D. J. Seo, J. A. Smith, "Rainfall estimation using raingages and radar. A Bayesian approach: 2. An application," *Stochastic Hydrology and Hydraulics*, vol. 5, pp. 31 - 44, 1991.
[17] U. Ehrel, *Rainfall and Flood Nowcasting in Small Catchments Using Weather Radar*, Ph.D. Thesis, University of Stuttgart, 2002.
[18] J. W. Wilson and E. A. Brandes, "Radar measurement of rainfall - a summary," *Bulletin of the American Meteorological Society*, vol. 60, no. 9, pp. 1048-1058, 1979.
[19] E. Habib and W. F. Krajewski, "Uncertainty analysis of the TRMM ground-validation radar-rainfall products: Application to the TEFLUN-B field campaign," *Journal of Applied Meteorology*, vol. 41, pp. 558-572, 2002.
[20] N. A. C. Cressie, *Statistics for Spatial Data*, Wiley, 1991.
[21] S. Sinclair and G. Pegram, "Combining radar and rain gauge rainfall estimations using conditional merging," *Atmospheric Science Letters*, vol. 6, issue 1, pp. 19-22, 2005.
[22] B. Schölkopf, P. L. Bartlett, A. J. Smola, and R. Williamson, "Shrinking the tube: a new support vector regression algorithm," *Advances in Neural Information Processing Systems*, vol. 11, pp. 330-336, 1999.

## 저 자 소 개

**유철상(Chulsang Yoo)**
1987년 고려대 토목공학과 졸업(학사)
1989년 고려대 대학원 토목공학과 졸업(석사)
1995년 Texas A&M University 토목공학과 졸업(박사)
2003년~현재 고려대 공과대학 건축사회환경공학과 교수

관심분야 : 수문학, 확률과정응용

Phone : 02-3290-3321
E-mail : envchul@korea.ac.kr

**박주영(Jooyoung Park)**
1983년 서울대 전기공학과 졸업(학사)
1985년 KAIST 졸업(석사)
1985년~1988년 한국전력 월성원자력발전소 근무
1992년 University of Texas at Austin 전기및컴퓨터공학과 졸업(박사)
1993년~현재 고려대 서창캠퍼스 제어계측공학과 교수

관심분야 : 계산지능, 서포트벡터학습, 강화학습

Phone : 041-860-1444
E-mail : parkj@korea.ac.kr