

Incremental Ensemble Learning for The Combination of Multiple Models of Locally Weighted Regression Using Genetic Algorithm

Kim Sang Hun[†] · Chung Byung Hee^{**} · Lee Gun Ho^{***}

ABSTRACT

The LWR (Locally Weighted Regression) model, which is traditionally a lazy learning model, is designed to obtain the solution of the prediction according to the input variable, the query point, and it is a kind of the regression equation in the short interval obtained as a result of the learning that gives a higher weight value closer to the query point. We study on an incremental ensemble learning approach for LWR, a form of lazy learning and memory-based learning. The proposed incremental ensemble learning method of LWR is to sequentially generate and integrate LWR models over time using a genetic algorithm to obtain a solution of a specific query point. The weaknesses of existing LWR models are that multiple LWR models can be generated based on the indicator function and data sample selection, and the quality of the predictions can also vary depending on this model. However, no research has been conducted to solve the problem of selection or combination of multiple LWR models. In this study, after generating the initial LWR model according to the indicator function and the sample data set, we iterate evolution learning process to obtain the proper indicator function and assess the LWR models applied to the other sample data sets to overcome the data set bias. We adopt Eager learning method to generate and store LWR model gradually when data is generated for all sections. In order to obtain a prediction solution at a specific point in time, an LWR model is generated based on newly generated data within a predetermined interval and then combined with existing LWR models in a section using a genetic algorithm. The proposed method shows better results than the method of selecting multiple LWR models using the simple average method. The results of this study are compared with the predicted results using multiple regression analysis by applying the real data such as the amount of traffic per hour in a specific area and hourly sales of a resting place of the highway, etc.

Keywords : Locally Weighted Regression, Multi Model Selection, Incremental Ensemble Learning, Genetic Algorithm

유전 알고리즘을 이용한 국소가중회귀의 다중모델 결합을 위한 점진적 앙상블 학습

김 상 훈[†] · 정 병 희^{**} · 이 건 호^{***}

요 약

전통적으로 나타낸 학습에 해당하는 국소가중회귀(LWR: Locally Weighted Regression)모델은 입력변수인 질의지점에 따라 예측의 해를 얻기 위해 일정구간 범위내의 학습 데이터를 대상으로 질의지점의 거리에 따라 가중값을 달리 부여하여 학습 한 결과로 얻은 짧은 구간내의 회귀식이다. 본 연구는 메모리 기반학습의 형태에 해당하는 LWR을 위한 점진적 앙상블 학습과정을 제안한다. LWR를 위한 본 연구의 점진적 앙상블 학습법은 유전알고리즘을 이용하여 시간에 따라 LWR모델들을 순차적으로 생성하고 통합하는 것이다. 기존의 LWR 한계는 인디케이터 함수와 학습 데이터의 선택에 따라 다중의 LWR모델이 생성될 수 있으며 이 모델에 따라 예측 해의 질도 달라질 수 있다. 하지만 다중의 LWR 모델의 선택이나 결합의 문제 해결을 위한 연구가 수행되지 않았다. 본 연구에서는 인디케이터 함수와 학습 데이터에 따라 초기 LWR 모델을 생성한 후 진화 학습 과정을 반복하여 적절한 인디케이터 함수를 선택하며 또한 다른 학습 데이터에 적용한 LWR 모델의 평가와 개선을 통하여 학습 데이터로 인한 편향을 극복하고자 한다. 모든 구간에 대해 데이터가 발생 되면 점진적으로 LWR모델을 생성하여 보관하는 열심학습(Eager learning)방식을 취하고 있다. 특정 시점에 예측의 해를 얻기 위해 일정구간 내에 신규로 발생된 데이터들을 기반으로 LWR모델을 생성한 후 유전자 알고리즘을 이용하여 구간 내의 기존 LWR모델들과 결합하는 방식이다. 제안하는 학습방법은 기존 단순평균법을 이용한 다중 LWR모델들의 선택방법 보다 적합도 평가에서 우수한 결과를 보여주고 있다. 특정지역의 시간 별 교통량, 고속도로 휴게소의 시간별 매출액 등의 실제 데이터를 적용하여 본 연구의 LWR에 의한 결과들의 연결된 패턴과 다중회귀분석을 이용한 예측결과를 비교하고 있다.

키워드 : 국소가중회귀분석, 다중 모델의 선택, 점진적 앙상블 학습, 유전알고리즘

※ 이 논문은 숭실대학교 교내연구지원비에 의하여 연구되었음.
† 정 회 원 : (재)한국화학융합시험연구원 선임연구원
** 비 회 원 : 숭실대학교 산업·정보시스템공학과 교수
*** 종신회원 : 숭실대학교 산업·정보시스템공학과 교수

Manuscript Received : January 10, 2018
First Revision : May 14, 2018
Accepted : June 25, 2018
* Corresponding Author : Gun Ho Lee(ghlee@ssu.ac.kr)

1. 서 론

전통적으로 메모리 기반 학습과정은 학습을 위한 데이터를 보존하고 예측이 필요할 때 데이터를 사용하여 예측 모델을 구축한다. 동시대의 대규모 데이터를 대상으로 하는 모델구축 학습과정에서는 예측을 위한 비모수화 모델을 구축하기 위해 학습을 한 후에는 데이터는 보존하지 않고 폐기하는 것에 많은 관심의 대상이 되고 있다. 메모리 기반 학습은 나태한 학습기법(lazy learning method)에 해당하며 문제 해결 요청이 있을 때 학습을 하고 예측 모델을 구축하여 질의에 대한 해를 도출한다.

국소가중회귀(LWR, Locally Weighted Regression)는 일종의 나태한 학습으로 학습 데이터의 평균, 데이터 간 보간 과정, 데이터들로부터 추론과정, 혹은 학습 데이터들을 병합하는데 LWR을 이용하는 것이다[1].

서포터 벡터 머신(Support Vector Machine), 인공신경망, 회귀분석과 같은 대부분의 전역모델은 문제 해결지점 외의 구간 데이터들도 학습의 대상이 된다. 하나의 문제를 해결하는 경우에도 모든 입력 데이터를 적용한 모델을 구축한 후 예측을 한다. 전역모델의 단점은 때때로 모수 값들이 세부적인 근사화를 시키지 못하며, 전역의 모든 데이터를 대상으로 학습하기 때문에 많은 시간이 소요되며 모델의 규모도 커질 수 있다. 하지만 K-최근접법, 가중평균기법, LWR 등의 지역모델(local model)은 문제해결 요청이 있는 질의의 위치 부근 데이터만을 학습하게 된다[2]. LWR의 기본적인 아이디어는 전 데이터 분포에 해당하는 회귀모델을 구축하는 대신 질의 지점에 이웃한 데이터들을 기반으로 국소 모델을 구축하여 해를 구한다. 따라서 LWR은 전역모델을 대체하여 학습시간을 단축하고, 모델 자체를 단순화시킬 수 있는 좋은 대안이 될 수 있다[2]. 전통적인 LWR은 많은 이점에도 불구하고 특정 질의에 대해 특정 구간 데이터만을 대상으로 한 모델이므로 노이즈나 비정상적인 상황의 데이터로 학습이 될 수 있다. 본 연구에서는 이러한 단점을 보완하기 위하여 과거의 패턴이나 경향을 기반으로 모델을 구축하는 점진적 앙상블 학습과정을 제안하고자 한다. 또한, 기존 LWR에서 인디케이터 함수의 선택과 샘플링으로 추출된 데이터 집합의 선택에 따라 해의 질이 좌우되는 어려움을 진화적 학습과정을 통해 해결하고자 한다.

이어지는 내용의 구성은 다음과 같다. LWR 관련연구와 수학적 기본모델을 2장에서 소개한다. 3장에서 일반적인 앙상블 학습에서 다중 모델의 선택방법에 대해 소개하며 4장에서는 데이터의 선택, 유전자 알고리즘을 이용한 해의 개선과정, 학습의 데이터 발생 상황을 포함한 본 연구의 앙상블 학습과정을 소개를 한다. 5장에서 일반적인 다중모델 결합방법인 단순 평균법과 본 연구 학습결과를 비교하고 다중회귀분석을 이용한 예측의 결과에 대한 적합도와 패턴을 본 연구의 결과를 연속하여 연결한 경우를 비교한다.

2. 국소가중회귀분석(LWR)

대규모의 데이터를 처리하는 환경에서는 선형회귀분석 혹은

다중회귀분석과 같은 전역모델보다 주변 데이터만을 분석하고 예측모델을 구축하는 국소모델이 더 유연하고 효율적인 결과를 도출하므로 다양한 분야에 적용되어 왔다. 주택 가격지수 추정[3], 고용 중심지 결정을 위한 후보지 식별[4], 차량 주행속도에서 보이는 변수 값들의 보정[5], 품질 예측의 학습과정에 이용하기도 했다[6]. 또한, 데이터가 특정 지역에 편중되는 현상을 보완하는 분석방법[7] 등에 대한 연구가 수행되어 왔다.

이동평균이 시계열에 의해 계산되는 것과 유사한 방식으로 LWR을 이용하여 예측 함수 값을 근사화 하고 이를 그래프로 나타내기 위해 산점도 평활방법을 적용하는 연구가 있었다[8]. 또한, LWR 적용 시 중요 인자인 윈도우 크기와 방향차수 값을 결정하기 위해 오차율을 최소화하는 방법이 제시되었다[9]. LWR을 이용하여 과거와 현재 자료를 활용한 예측모형에 대한 연구도 수행 되었다[10].

Meier 외 2인은 점진적 국소가우시안 회귀법을 제시한 바 있다[11]. 이는 국소 모델을 찾아내기 위해 전체 범위의 모델의 목적함수(global objective)를 재 공식화를 수행하는 탑-다운(Top-down) 방식을 적용하며 LWR과 유사한 국소회귀 알고리즘은 전체 범위를 대상으로 하는 모델을 근사화와 수정과정을 병행한다. 국소가우시안 회귀법은 다수의 국소 비모수적 가우시안 프로세스 모델들을 가진 예측 모델형태로 쉽게 확장이 가능하다는 특징이 있다. Talgorn외 3인은 일명 근사치 최적화 방식인 썬로게이트(surrogate) 최적화를 위한 국소가중산점도 평활(LOcally WEighted Scatterplot Smoothing; LOWESS) 모델의 효과적인 사용을 위해 3가지 제안을 하고 있다[12]. LOWESS모델 계산비용을 줄이는 방법과, 평활을 유지하는 동안 인접한 데이터의 분포에 LOWESS모델들을 적용하기 위해 스케일링(scaling)계수를 도입하고 있다. 또한 LOWESS 모델의 최적 형상 계수를 선택하기 위하여 적절한 차수 오차 측정법을 사용한다. 썬로게이트 최적화 방법은 메쉬 적응적 직접 탐색 알고리즘(Mesh Adaptive Direct Search; MADS)을 사용하고 있다. MADS알고리즘에서는 유망한 후보를 생성하고 순위를 매기는데 LOWESS 모델들이 사용된다.

과거 데이터를 이용하여 회귀분석 시 회귀선의 형태 및 오차의 분포에 대한 가정 하에 회귀식을 유도한다. 만약 데이터가 가정에 부합되지 않으면 유도된 회귀식을 신뢰하기 어려울 것이다. 데이터에 대한 지식이 없는 경우에는 기존의 전통적인 선형회귀분석 방법보다는 탐색적 접근 방법에 의한 비모수적 회귀분석이 더 유용할 수 있다. 이 방법은 회귀선의 형태를 직선으로 국한시키지 않으며, 자료의 구조에 대한 전반적인 통찰력을 얻는데 목적을 두고 있다[13]. 이처럼 비모수적 회귀분석은 모수적 방법에 의한 전통적인 회귀분석이 실제 자료에서 흔히 나타나는 기복현상을 표현하는데 적절하지 않기 때문에 이에 대한 대안으로 관심의 대상이 되고 있다[4]. 비모수적 회귀함수 추정 방법에는 스플라인 평활화, k-최근접, 커널추정, LWR 등이 있다[13]. 이 중 LWR은 예측을 요구하는 질의지점에 가까운 데이터를 중심으로 인접한 데이터일수록 큰 가중치를 부여하여 한다. 이웃한 데이터들은 데이터 간 거리 차에 따라 가중치를 부여하며, 이때 사용

하는 함수로는 가우시안함수, 트리큐빅(tricubic)함수, 지수함수, 2차 함수(quadratic) 등이 있다[6].

인공신경망과 가우시안 혼합모델과 같은 모델 기반의 방법들은 모델을 만드는데 사용한 데이터는 학습 후 폐기한다. 메모리 기반의 방법들은 학습 데이터를 유지하는 비모수적 접근방법을 따르며, 매번 예측이 필요한 시점에 사용한다.

LWR 모델 학습을 위한 기본적인 아이디어는 Equation (1)~(3)으로 간단히 나타낼 수 있다. LWR은 선형함수를 사용하는 지역 주변의 질의가 들어온 시점에서 목표함수를 근사화할 수 있으며, 목표함수 f 를 추정하기 위해 Equation (1)의 선형함수를 사용한다. w_0, \dots, w_n 은 주어진 학습 데이터에 대하여 선형함수 평활화 과정에서 Equation (2)의 에러 값을 최소화하기 위한 계수이다. Equation (1)의 함수값과 목표함수의 차이인 잔차를 최소화하기 위해 반복 수행하게 된다. Equation (2)의 총 에러 값의 제곱(total squared error)값을 E 로 정의(≡)하며, 이를 최소화하기 위해 Equation (3)의 가중치 함수를 적용한다.

$$\hat{f}(x) = w_0 + w_1 a_1(x) + \dots + w_n a_n(x) \quad (1)$$

$$E \equiv \frac{1}{2} \sum_{x \in D} (f(x) - \hat{f}(x))^2 \quad (2)$$

$$\Delta w_j = \eta \sum_{x \in D} (f(x) - \hat{f}(x)) a_j(x) \quad (3)$$

여기서 $f(x)$ 는 목표함수(target function), $\hat{f}(x)$ 는 추정 함수값, D 는 데이터 집합 $D = \langle x, f(x) \rangle$, η 는 학습률을 각각 나타낸다. 데이터 x 는 n 개의 속성을 가지며, $x = (a_1(x), a_2(x), \dots, a_n(x))$, $a_j(x)$ 는 데이터 x 의 j 번째 속성의 값을 나타낸다. 데이터 집합 D 를 단순화하기 위해 질의지점 x_q 에 가장 가까운 k 개의 데이터를 대상으로 학습을 수행할 수도 있다.

LWR은 가까이 위치한 데이터를 이용하여 평활화시키는 각 포인트의 다항식을 추정하는 방법으로 자료의 선형성, 주기성에 따라 발생하는 여러 시계열 데이터에 적용이 가능하다[9].

LWR의 장점을 5가지로 정리해볼 수 있다.

첫째, 알고자 하는 시점을 기준으로 이전과 이후의 데이터를 모두 참조하여 평활화를 수행한다. 둘째, 가중치 함수에 따라 다양한 값을 부여할 수 있다. 셋째, 평활화 또는 예측하기 위해 참조하는 인접 데이터의 개수를 결정하는 문제해결 대상 구간 t 와 함수식의 차수를 결정하는 다항식의 차수 값(polynomial order)에 따라 평활정도를 조정할 수 있다. 넷째, 추정되는 결과 값의 범위를 설정할 수 있어 분석 시 다양한 제한조건을 적용할 수 있다. 마지막으로 분석 대상 지점을 기준으로 전후에 인접한 자료에 가중치를 부여하고 최소자승법을 통해 함수식을 구성한 후, 추정한 결과 값을 산출할 수 있다[5].

선형회귀보다 유연하고 학습능력이 좋아 패턴분석에 좋은 효과를 보여주고 있음에도 LWR을 수행하는 데 있어 몇 가지 극복해야 할 어려움이 있다. 첫째, 인디케이터(indicator function) 혹은 커널(kernel) 함수에 따라 모델이 달라질 수 있다. 둘째, 데이

터에 적합한 구간을 찾아야 한다. 셋째, 데이터 샘플의 선택에 따라 결과가 달라질 수 있다. 본 연구에서는 이러한 문제를 해결하기 위해 적절한 인디케이터 함수의 선택과 데이터 샘플에 따른 모델을 구축하는 점진적 앙상블 학습절차를 제시하고자 한다.

3. 다중 모델의 선택방법

앙상블 학습법은 다수 예측 모델을 구축하고 그 중에서 우수한 예측 기능들을 결합하는 것이다[14, 15]. 따라서 단일 예측 모형보다 좋은 예측 값을 얻을 수 있다는 장점이 있다.

다중 모델을 생성과 이들의 결합의 방법으로 데이터 조정법, 가설 열거법, 무작위 인제팅법, 입력 속성조정법, 출력 목표 값의 조정법이 있다[16]. 일반적으로 다수의 모델(가설)들을 생성하기 위해 학습 알고리즘을 수차례 실행 할 때 마다 전체 학습 데이터 중 일부분을 적용하는 데이터의 조정방법이 널리 사용되고 있다. 이는 학습 데이터의 작은 변화에 대해 결과의 값이 크게 변화하는 불안정한(unstable) 인공신경망, 규칙학습 알고리즘, 트리생성 알고리즘에 적합하다[16].

K-fold 교차검증, 배깅, 부스팅, 스택킹, 아다부스트(AdaBoost) 등이 대표적인 데이터 조정을 이용하는 앙상블 학습이다. 이들은 Table 1과 같이 학습 데이터의 배분방식 뿐만 아니라, 다중 모델들의 선택 방식, 수행척도와 이를 위한 사용 알고리즘에 따라 특성을 달리하고 있다[17].

일반적으로 사용되는 간단한 다중 모델들의 선택 방법에는 투표방법과 평균방법이 있다. 결과 값의 타입이 범주형인 경우는 투표법을 적용하며 결과 값이 수치형인 경우는 평균법이 적용된다. 투표방법 중 다득표 투표법은 각 테스트 데이터에 대하여 모든 모델은 예측을 수행하고 예측한 결과에 해당하는 클래스가 전체 결과 수의 1/2 수를 넘으면 이에 해당하는 클래스를 최종 결과로 하는 것이다. 만약 모델 예측의 어느 결과도 1/2 수를 넘지 못하면 안전한 예측을 할 수 없는 것이다. 이때 각 모델은 동등한 비중을 갖는 민주적인 투표방식이다. 하지만 가중치 투표법(Weighted Voting)은 다 득표 투표법과 달리 중요한 모델일수록 투표결과에 높은 값을 부여하는 방식이다.

Table 1. Comparison of Combiners

Combiner Characteristic	Bagging	Boosting	Stacking
Partitioning of the data into subsets	Random	Giving mis-classified higher preference	Various
Goal to achieve	Minimize variance	Increase predictive force	Both
Methods used	Random subspace	Gradient descent	Blending
Function to combine single models	(Weighted) average	Weighted majority vote	Logistic regression

평균법에도 단순 평균법(Simple Averaging)과 가중 평균법(Weighted Averaging)이 있다. 단순평균법은 테스트 데이터를 예측 모델에 적용하여 예측 값들의 평균을 최종 예측 값으로 한다. 이는 과적합(overfitting)을 줄여 줄 수 있으며 매끄러운 회귀선을 생성 할 수 있는 장점이 있다. 가중 평균법도 단순평균법과 비슷하며 각 모델의 과거 예측 성적이 우수한 모델이거나 중요한 모델에 높은 가중값을 부여하여 가중평균값을 구하여 최종 예측 값으로 한다.

4. 점진적 앙상블 학습과정의 제안

본 연구에서는 Fig. 1과 같이 다수의 예측 모델들을 생성하고 결합하는 과정을 유전알고리즘을 이용하여 단일 분석 모델보다는 좋은 예측 값을 갖는 모델을 얻고자 한다. 실시간으로 축적되는 데이터에 대응하기 위해 특정구간 내에서 무작위 샘플링을 한 데이터를 대상으로 LWR을 수행하여 모델을 생성하며, 이 과정을 반복하여 다수의 분석 모델을 생성할 수 있다. 생성된 모델들을 결합하고 유전알고리즘을 통해 충분히 진화시켜 가장 좋은 모델을 도출한다. 랜덤 샘플링을 통해 표본데이터를 추출하고, 분석 단계에서 적절한 인디케이터 함수를 선택할 수 있다. 또한, 유전알고리즘을 통하여 복수의 함수를 대상으로 점진적 앙상블 학습을 수행하여 최적의 모델을 구축하고자 한다.

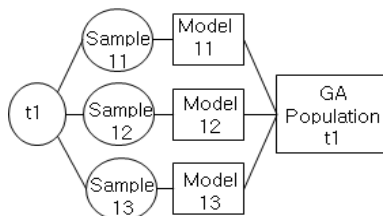


Fig. 1. Ensemble Learning

본 연구에서는 유전알고리즘을 이용한 LWR기반 점진적 앙상블 학습절차를 제시한다. 여기서 t 는 데이터의 발생 시점, d 는 데이터 구간의 크기, K 는 데이터 수를 나타낸다.

- Step 1: 시점 t 를 정하고, t 를 중심으로 $2d$ 구간 내의 데이터를 수집한다.
- Step 2: 수집된 데이터에서 랜덤 샘플링을 통해 표본 데이터를 K 개 추출한다.
- Step 3: 추출된 K 개의 표본 데이터에 LWR을 적용하고, 생성된 LWR 모델들은 초기 해집단이 된다.
- Step 4: t 시점 이후 새로운 데이터가 발생하면, 이를 기반으로 Step 1~3을 반복 수행한다. 기존 Step 3에서 얻은 해집단은 수행 후 얻어진 해집단과 결합하여 새로운 해집단을 생성한다.
- Step 5: 새로운 데이터의 입력이 없거나 사용자가 지정한 시점이 되면 전 단계의 해집단을 기반으로 유전알고리즘을 수행하여 모델을 구축한다.

전체 데이터를 분할하여 표본을 추출하는 것이 이상적이지만 이를 대신할 데이터의 표본을 무작위 추출한다. Step 3에서는 Step 2에서 추출된 K 개의 표본 데이터에 LWR을 적용하여 모델을 구축하여 이들을 유전알고리즘의 해집단에 포함시킨다. Step 4에서 해집단을 구축 후 새로운 데이터가 생성되는 상황이라면 추가적인 데이터들에 대한 분석을 위해서 새로운 해집단을 만들어야 한다. 점진적 앙상블 기법을 응용하여 새로운 데이터에 대한 해집단을 구하고, 이미 분석된 해집단과의 결합을 통해 분석을 위한 새로운 해집단을 만들어 낸다. Step 5에서 새로운 데이터가 발생되지 않거나, 사용자가 지정한 시점에 도달하면, Step 4에서 구한 해집단을 기반으로 유전알고리즘을 수행하여 최적모델을 추출한다.

Step 1~2의 데이터 수집과 추출은 4.1, Step 3의 LWR 적용과 초기 모델 구축은 4.2, Step 4~5의 유전알고리즘은 4.3에서 세부적인 설명을 한다.

4.1 데이터 샘플링

대규모의 데이터를 직접 처리하는 것은 많은 시간과 비용이 소요되므로 샘플링 한 데이터를 대상으로 한다. 샘플링은 시점 t 를 중심으로 $2d$ 의 구간 내에서 S 등분으로 분할한다. S 개의 구간에서 균등하게 샘플링하여 총 K 개 데이터를 추출한다.

샘플링의 특성상 샘플에 따라 모델의 패턴이 달라지며 그 예측율도 다르게 나타날 수 있다. 이러한 부분을 극복하기 위하여 다수의 추출 데이터 집합을 확보한 후 점진적 앙상블 학습과정을 반복 수행한다.

4.2 국소가중회귀분석 적용

LWR 모델 구축 시 가중값 부여를 위해 Rectangular함수, Tricube함수, Triangular함수, Epanechnikov함수, Quartic함수, Triweight함수, Gaussian함수, Cosine함수, Logistic함수 등이 인디케이터 함수로 사용된다. 일반적으로 Tricube함수와 Gaussian함수를 사용하지만 인디케이터 함수에 따라 예측 능력은 다를 수 있다[6]. 분석상황에 따라 어떤 인디케이터 함수가 적절한지를 선택하는 것은 어려운 일이다. 본 연구의 학습절차에서는 인디케이터 함수들을 다양하게 적용하여 가장 적합한 함수를 선택하거나 순차적으로 적용할 수 있도록 한다. LWR모델 구축단계에서는 7개의 인디케이터함수를 적용한다. 데이터를 LWR 알고리즘에 적용 시 윈도우 값은 0.25로 고정하여 인디케이터 함수들을 순차적으로 적용하여 비교평가 한다.

LWR의 기본적인 함수의 형태는 Equation (4)와 같다. t 는 예측 시점, t_0 는 추정 시점, K 는 데이터의 개수, β_0 는 t 의 추정된 함수인 다항식의 계수이다.

t 의 값을 예측하기 위해 Fig. 2와 같이 t_0 와 t_0 에 인접한 K 개의 데이터를 사용한다. Equation (5)는 t_0 의 데이터와 가중치를 부여한 인접한 K 개의 데이터를 최소자승법을 기반으로 하여 다항함수를 추정하는 식이며 가중치 행렬(W_{10})은 대칭행렬이다[9].

$$y(t) = f_{t_0}(t, \beta_{t_0}) + \epsilon_{t_0,t} \quad (4)$$

$f_{t_0}(t, \beta_{t_0})$: 추정 국소가중회귀함수에 의해 시점 t_0 에서

예측한 시점 t 의 값

$\epsilon_{t_0,t}$: 정규분포를 가지는 잔차

X_{t_0} : 시점 t_0 를 중심으로 인접한 데이터 행렬

W_{t_0} : 인접 데이터에 부여되는 가중치 행렬

$$\min_{\beta_{t_0}} [X_{t_0} - f_{t_0}(t, \beta_{t_0})]^T W_{t_0} [X_{t_0} - f_{t_0}(t, \beta_{t_0})] \quad (5)$$

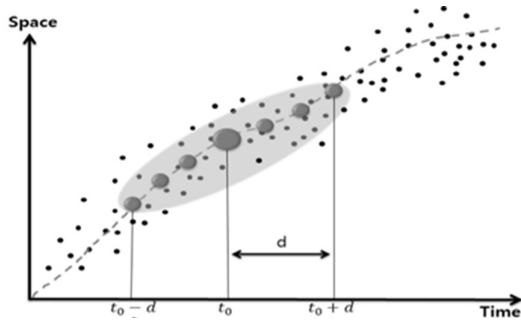


Fig. 2. Distribution of Data Over Time

LWR을 수행 시 추정되는 함수의 차수는 다항식의 차수 값에 의해 결정되며 이에 따라 추정계수(β)의 개수도 달라진다. 함수 추정 시 사용되는 시점 t_0 를 기준으로 인접한 데이터의 개수를 문제해결 대상 구간 t 의 길이만큼 설정할 수 있으며, t 를 크게 하는 것은 추정 시 참조할 수 있는 주변 데이터의 수를 많이 설정하므로 자료를 평활화하는 정도가 증가한다[9].

LWR 적용 후 시간과 보정값 등의 결과 값을 유전알고리즘의 진화과정을 위해 초기 해집단에 저장한다.

4.3 유전알고리즘을 이용한 해의 개선

본 연구에서는 추출한 표본 데이터를 대상으로 LWR알고리즘을 적용하여 다양한 예측모델을 생성한 후, 가장 높은 예측을 혹은 정확도를 갖는 최적 모델을 도출해내기 위해 유전알고리즘을 이용한다. 점진적 앙상블기법을 응용하여 데이터들을 모델화하여 유전알고리즘의 해집단으로 입력된 후에는 사용자 전략에 따라 평가함수, 선택, 교배, 돌연변이의 연산을 수행할 수 있다. 유전알고리즘 수행 시 세부 사항에 대해서는 사용자의 목적에 맞게 설정한다.

본 연구에서 사용한 평가함수는 저장된 모델이 가지고 있는 추정치(fitted value)와 잔차(residual value)를 이용한다. 추정치는 학습 데이터에서 가중치가 적용되어 수정된 값을 의미하며, 잔차는 LWR이 적용된 후 학습 데이터의 값과 추정치의 차이를 의미한다. Equation (6)에서 잔차가 작을수록 실제 학습 데이터와 추정치가 유사함을 의미한다.

$$fitness = \sum_{x \in D} (f(x) - \hat{f}(x))^2 \quad (6)$$

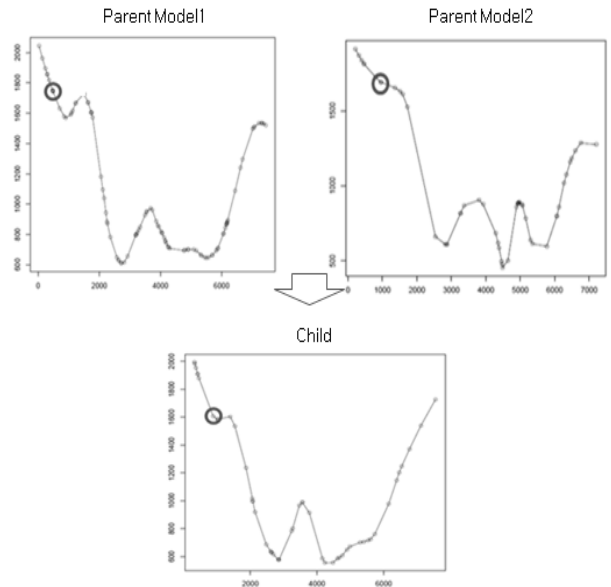


Fig. 3. Crossover of Two Models

유전 알고리즘을 적용한 모델의 개선 및 통합 과정은 다음과 같다.

1) 초기 LWR 모델을 해집단의 모델로 사용

2) 선택 단계

선택연산은 기본적으로 룰렛-휠(roulette wheel)에 의해 수행한다. 부모 중 우수성을 평가하기 위해 Equation (6)의 평가함수를 적용한다. 해집단의 다양성 보존과 조기 수렴을 방지하기 위해 부모 해 중 하위 50%에 대해서도 교배를 위한 한 부모를 선택하도록 한다.

3) 교배 단계

교배 후보자는 선택 단계의 평가방식에 따라 교배 후보자를 선택한다. 교배는 Fig. 3과 같이 선택된 부모 해의 각 속성값의 산술평균 즉 $x_c=(x_1+x_2)/2$, $y_c=(y_1+y_2)/2$ 형태로 자식 해를 생성한다.

4) 변이 단계

돌연변이는 0.3의 확률로 변이 대상의 인자를 선택하게 한다. 교배 이후 변이를 수행하여 m 개의 자식 해를 생성한다. 교배단계에서는 3.5m개 중 2개 부모 해를 선택하여 $x_c=(x_1+x_2)/2$, $y_c=(y_1+y_2)/2$ 형태로 자식 해를 생성한다. 이 과정을 거치면 총 $3.5mC_m$ 개의 자식 해가 생성되고, 이에 대한 변이과정을 진행한다. 추가로 변이에 선택된 염색체에 대해서는 설정한 변이함수를 통해 변이를 시행한다. 설정한 변이함수는 선택된 염색체를 50% 확률로 $(1 \pm \text{가중치})$ 만큼 변이를 주는 것이다. 실행 후 새롭게 생성된 자식 수는 $3.5m(3.5m-1)/2$ 가 되고, 이는 부모 해를 대체하는 데 필요한 자식 해 보다 더 많은 수이다.

따라서 문제해결 대상 구간에서 $3.5m(3.5m-1)/2$ 개 중 평가함수 값이 우수한 자식 해 3.5m 개를 선택한다. 유전알고리즘 선택, 변이과정은 사용자 지정만큼 수행할 수 있다.

5) 병합 단계

변이단계를 통해 분할구간별 최종적인 해집단이 도출되면 이를 통합하는 과정이 필요하다. 연속적 혹은 간헐적으로 데이터가 점진적으로 발생하게 되면 과거에 구축한 모델과 현재 데이터를 기반으로 구축한 모델을 결합한다. 또한, 현재와 과거 시점의 데이터도 추출하여 사용하므로 일시적인 중복이 존재할 수 있으며 이 경우, 가장 최신의 것을 먼저 선택하도록 한다. 새로운 지식 해들을 대상으로 변이율만큼 변이를 진행하며, 선택된 염색체에 대해서는 설정한 변이함수를 통해 변이를 수행한다.

6) LWR의 비교

K개의 학습 데이터 집합에 대하여 2)~5)의 과정의 반복을 거쳐 도출한 최종 해집단에 7개의 인디케이터 함수를 모두 적용하여 LWR의 비교분석을 진행한다. K개 학습데이터 집합에 각각 인디케이터 함수를 적용한다. 적용된 함수의 LWR를 LWR 1~7까지로 표현하면 초기 해집단에서 이 과정을 거쳐 나온 해집단의 개수는 7K 개가 되며, 유전알고리즘의 해집단으로 입력되어 선택, 교차, 변이 과정이 반복된다.

4.4 점진적 앙상블 학습과정의 상황

시간의 변화에 따라 점진적으로 발생하는 대규모의 데이터를 일시에 일괄적으로 분석하면 시간과 노력이 많이 소요된다. 특히, 일정 시간 지속 후에는 새로운 데이터와 과거에 분석한 데이터들이 다시 포함되어 더 많은 시간과 노력이 필요할 것이다. 이를 피하고자 본 연구에서는 데이터가 발생에 따라 앙상블기법에 점진적 학습방법을 접목한 점진적 앙상블 학습과정을 제시하고자 한다. 학습 알고리즘은 다양한 상황에서도 제약 없이 데이터의 적용할 수 있어야 할 것이다. 본 연구에서는 데이터의 발생이 연속적이거나, 간헐적인 상황 모두 점진적 앙상블 기법을 이용한 해집단화가 가능한 학습과정을 제시하고 있다.

1) 연속적 데이터 발생 상황

연속적 데이터 발생 상황에서 점진적 학습방법은 현재 주어진 데이터로부터 그 결과를 생성하고 연속하여 새로운 데이터가 발생 시 이를 반영하여 현재의 모델을 수정하는 학습 방법이다. Fig. 4와 같이 t1 이전 시점에서 앙상블기법을 응용하여 데이터 샘플 집합별로 LWR모형을 얻게 된다. 이 모델들은 t1 시점에서 새로운 모델을 구축할 때 유전알고리즘의 초기 해집단으로 사용된다. t2 시점에서는 t1 시점에서 구축한 모델들이 t2 시점에서 준비된 해집단과 함께 유전알고리즘 초기 해집단으로 사용되어 새로운 모델을 구축하게 된다. t1 이후 t2 시점까지 데이터에 대해서 t1 시점에서 수행한 학습과 같이 수행한다. 이 경우, 최근 발생한 데이터가 과거 데이터 보다 LWR모형에 점진적으로 더 적극적으로 반영되도록 한다. 긴 시간에 대한 모델을 얻으려면 t2 시점의 유전알고리즘 해집단은 t1시점의 해집단과 결합하여 최적화를 진행한다. 최적화 진행 후 얻은 모형은 [t1, t2] 구간을 아우르는 예측모형이 된다.

본 연구의 제안 내용은 분석된 데이터를 사용하여 질의 시점별로 유전알고리즘의 해집단을 구성하여 보관하면 과거 분석된 데이터들에 대해서는 반복적으로 분석을 진행할 필요가 없다는 것이다. 이미 분석된 해집단에서 좋은 모델을 추출하여, 데이터 발생 시점에서 분석된 해집단과 결합을 통해 전체 분석결과를 도출한다.

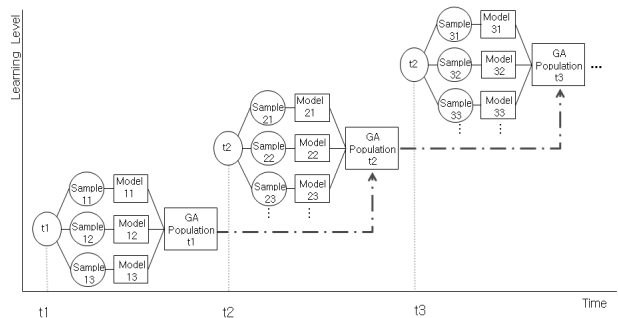


Fig. 4. Incremental Ensemble Learning of Continuous Situations

2) 간헐적 데이터 발생 상황

데이터가 비연속적이고 산발적으로 발생하는 상황에도 제시한 학습과정을 Fig. 5와 같이 적용할 수 있다. 수행과정은 연속적 앙상블 학습과 동일하다. Fig. 5에서와 같이 예를 들어 2015년 2월(s1), 2016년 2월(s2) 당시의 데이터가 이미 분석되어 해집단으로 보관 중이고 2017년 2월(s3)에 데이터 분석을 하는 경우, 2015년과 2016년 해당 시점의 해집단으로부터 좋은 모델을 추출하여 2017년 2월 데이터를 분석하는 유전알고리즘 해집단에 추가한다.

과거의 해집단들과 현시점의 분석을 통해 얻은 새로운 해집단과의 결합을 통해 분석결과와 질을 높일 수 있다. 이 방법은 과거의 분석된 데이터들이 존재하는 경우 비슷한 시점의 데이터 패턴의 예측에 유용하게 적용될 수 있다.

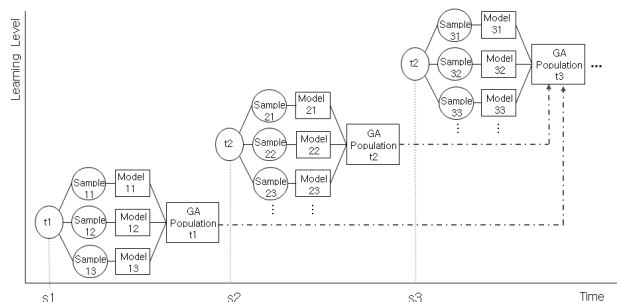


Fig. 5. Intermittent Incremental Ensemble Learning

5. 학습결과 및 분석

본 연구의 학습절차는 R언어로 구현하였다. 수행한 컴퓨터는 윈도우 64bit 운영체제와 CPU는 Intel i5 프로세서, Ram 4GB의 크기를 사용한다.

단순 평균법과 본 연구를 비교하기 위하여 8개의 데이터

셋을 균등 분포의 데이터 표본에서 무작위로 추출하여 적용하였다. Fig. 6에서 각 문제에 대해 10개의 예측 값의 결과와 평균값을 본 연구결과의 적합도 값들을 보여주고 있다. 4개의 문제에서 본 연구의 결과가 단순 평균법 보다 우수한 결과를 보여 주고 있으며 나머지는 비슷한 결과를 볼 수 있다.

본 연구에서는 Step 1에서 적절한 데이터 구간을 찾기 위해 전체 데이터 분포의 5등분을 단위로 앙상블학습과정을 적용했다. Fig. 7에서는 유전알고리즘 해집단의 크기를 60으로 설정하여 CPU시간 변화를 나타내고 있다. 데이터 구간 25를 기점으로 CPU시간은 급격하게 증가하는 것을 알 수 있으며, 효율적인 학습을 수행하기 위해 분할구간은 25로 설정하였다.

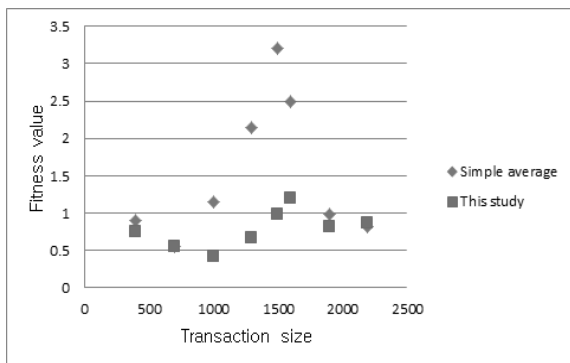


Fig. 6. Comparison of Fitness Value

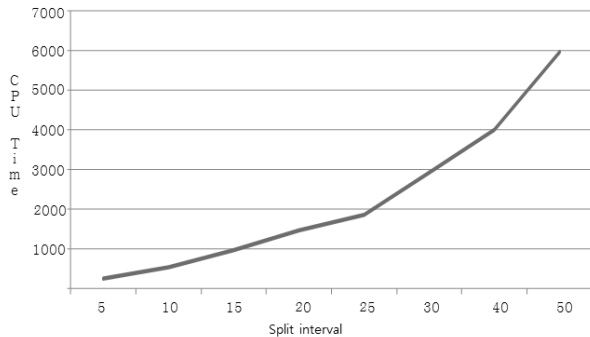


Fig. 7. CPU Time(sec.) Variation According to Data Interval

앙상블 학습절차 Step 3에서는 유전알고리즘의 적절한 해집단의 크기를 찾기 위해 다수의 해집단 크기별로 학습을 수행하였다. Fig. 8은 해집단의 크기와 CPU시간 관계를 나타내고 있다. CPU시간은 110초 범위 내에서는 크기에 거의 비례하여 증가하므로 110초 이내로 제한하였다. 또한, 학습절차의 수행에 필요한 데이터의 증가에 따른 알고리즘의 CPU시간을 조사하기 위해 균등분포를 이루고 있는 학습데이터를 무작위로 추출하여 사용하였다. Fig. 9에서와 같이 데이터의 증가에도 합당한 시간 이내에 학습을 수행할 수 있었다.

LWR의 점진적 앙상블 학습과 직접 비교할 수 있는 연구의 대상이 없으므로 동일 구간을 대상으로 다중회귀분석을 수행한 결과와 비교하기 위하여 대상 구간 t에서 일정 간격으로 점진적 앙상블 LWR을 수행한 다음, 다수의 패턴을 연결한 결과를 비

교하고자 한다. 부록에 첨부된 Table 2에서 문제의 데이터는 연령과 성적, 휴게소의 시간별 매출, 특정 지역의 시간별 교통량, 특정 기간의 보험청구액 등에 대한 것이다.

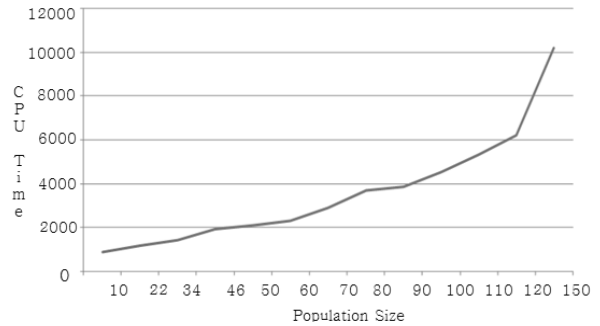


Fig. 8. The Population Size and CPU Time(sec.) Variation

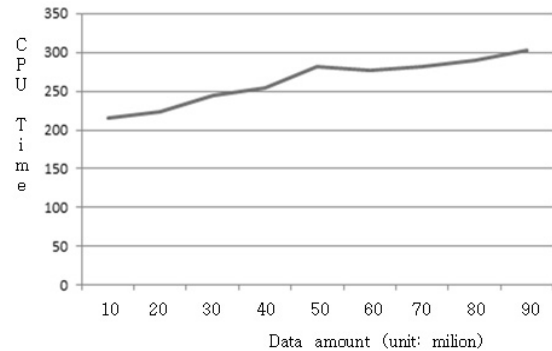


Fig. 9. CPU Time(sec.) with Data Amount

다중회귀법의 적용결과와 본 연구의 결과를 비교하기 위해 5개 문제에 해당하는 테스트 데이터 셋을 준비하여 실제 값과 학습한 모델이 예측한 결과에 대한 잔차의 제곱에 해당하는 적합도 Equation (6)에 적용하였다.

테스터를 위한 데이터 셋의 질의지점은 무작위로 선정하였고 모델의 각 데이터 셋의 적합도는 Table 2에 정리하였다. 본 연구가 적합도에서 비교적 우수한 결과를 나타내고 있다.

잔차의 제곱 값에 해당하는 값이 0에 수렴할수록 정확한 예측이라고 볼 수 있다. 문제 1의 경우 유전알고리즘의 특성상 진화를 위한 알고리즘 반복회수의 설정규칙을 정교하게 하면 잔차의 값은 더욱 개선될 수 있을 것으로 본다. 전반적으로 다중회귀분석은 짧은 구간보다는 긴 구간에 대응하는 완만한 패턴을 보인다. 반면에 본 연구의 패턴은 짧은 구간 내에서 비교적 변화가 두드러지는 것도 볼 수 있다. 어떤 연구가 절대적으로 우위에 있다고 판단할 수 있는 근거는 없다. CPU수행시간 또한, Fig. 9와 비슷한 경향을 보여주고 있다.

6. 결론

전통적인 LWR기법에서는 필요에 따라 특정 구간에 한해서 부분적인 회귀분석을 수행하고 사용한 데이터들은 폐기하

곤 한다. 이 경우 새롭게 분석하는 특정 구간의 데이터들이 비정상적이거나 일시적인 특이현상을 보이면 분석한 예측모형 역시 비정상적이거나 일시적인 패턴이 될 수 있다. 이러한 문제점을 보완하기 위해 본 연구에서는 과거의 경향과 패턴도 부분적으로 반영하고 있다. 과거 데이터 대신 과거 예측모형을 기반으로 현재 분석대상의 데이터와 새로운 학습을 하도록 하였다.

또한, 일반적인 규칙, 트리, 인공신경망, 서포트 벡터머신, 인공신경망 등의 예측 모델은 일정 기간 내에 발생하는 데이터만을 수집하여 일회성 모델을 구축하므로 지속적으로 발생하거나 산발적으로 발생하는 새로운 데이터에 대해 반응을 하기 어렵다. 본 연구에서 제시한 학습절차는 데이터가 연속적으로 발생하는 상황뿐만 아니라 간헐적으로 발생하는 상황에서도 데이터의 발생에 따라 모델의 갱신이 가능하도록 하였다.

본 연구는 과거 데이터 대신 이를 기반으로 구축한 모형을 현 학습단계에서 반영하여 단기간뿐만 아니라 장기간의 경향도 반영되도록 하고 있으며, 과거 데이터의 저장을 위한 대용량의 데이터베이스 없이도 이를 이용하는 것과 동등한 효과를 볼 수 있다. 또한, 제안하는 LWR을 위한 앙상블 학습과 점진적 학습모델의 통합은 기존 LWR에서 커널함수에 따라 해의 질이 결정되는 단점과 데이터의 샘플에 따라 구축된 모델의 예측능력의 차이를 극복할 수 있다는 것에 본 연구의 의의가 있다.

점진적 앙상블 학습을 위해 유전알고리즘의 1세대를 처리 시간이 짧고 반복학습과정이 단순하도록 데이터를 샘플링하여 모델을 구축한다면 실시간 형태의 시스템에도 사용 가능하며, 모델기반 추론시스템, 지식기반 추론시스템, 사례기반 추론시스템에도 유용하게 적용될 수 있을 것이다.

References

[1] V. Vapnik and L. Bottou, "All: local algorithms for pattern recognition and dependencies estimation," *Neural Computation*, Vol.5, No.6, pp.893-909, 1993.

[2] C. Atkeson, A. Moore, and S. Schaal, "Locally weighted learning. *Artificial Intelligence Review*," Vol.11, No.1-5, pp.11-73, 1997.

[3] H. Park, "A Study on the Construction of the Transaction-Based Real Estate Price Index Using Locally Weighted Regression(LWR) Model," *Journal of the Korea Real Estate Analysis Association*, Vol.17, No.1, pp.55-66, 2011.

[4] M.-J. Jun, "Identification of Seoul's Employment Centers by Using Nonparametric Methods," *Journal of Korea Planning Association*, Vol.39, No.3, pp.69-83, 2003.

[5] H. S. Lim, C. Oh, J. H. Park, and G. Q. Lee, "Study on Individual Vehicle Traveling Speed Filtering Method Using Locally Weighted Regression (LWR)," *Journal of Korean Society of Transportation*, Vol.59, pp.1094-1102, 2008.

[6] J. K. Cho, D. E. Lee, S. O. Song, and E. S. Yoon, "Quality estimation Using Support Vector Machine based on locally

weighted regression," *Journal of the Korean Institute of Gas*, Vol.10, pp.126-130, 2003.

[7] J. Kim, J. Lee, S. Y. Kim, and B. H. Lee, "The Effects of Point Accumulation Effort Level on Redemption Behavior in Loyalty Program," *Journal of Korean Marketing Association*, Vol.27, pp.85-106 2012.

[8] W. S. Cleveland, "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, Vol.74 No.368 pp.829-836, 1979.

[9] T. Toledo, H. Koutsopoulos, and K. Ahmed, "Estimation of vehicle trajectories with locally weighted regression," *Journal of Transportation Research Board*, Vol.1999, pp.161-169, 2007.

[10] H. Sun, H. Liu, H. xiao, R. He, and B. Ran, "Use of Local Linear Regression Model for Short-Term Traffic Forecasting," *Journal of Transportation Research Board*, Vol.1836, pp.143-150, 2003.

[11] F. Meier, P. Hennig, and S. Schaal, "Incremental local Gaussian regression," in *Proceedings of Advances in Neural Information Processing Systems, Montreal*, Vol.27, 2014.

[12] B. Talgorn, C. Audet, M. Kokkolaras, and S. Le Digabel, "Locally weighted regression models for surrogate-assisted design optimization," *Optimization and Engineering*, Vol.19, Issue 1, pp.213-238, 2018.

[13] S. Lee, "A Study on the Locally Weighted Regression," *Journal of Applied Science*, Vol.7, No.1, pp.121-129, 1998.

[14] R. Polikar, S. Krause, and L. Burd, "Ensemble of Classifiers Based Incremental Learning with Dynamic Voting Weight Update," *IEEE Xplore*, Vol.4, pp.2770-2775, 2003.

[15] Z. Erdem, R. Polikar, F. Gurgen, and N. Yumusak, "Ensemble of SVMs for Incremental Learning," In: Oza N.C., Polikar R., Kittler J., Roli F. (eds) *Multiple Classifier Systems. MCS 2005. Lecture Notes in Computer Science*, Vol.3541, pp.246-256, 2005.

[16] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Multiple Classifier Systems, Lecture Notes in Computer Science*, Vol.1857, pp.1-15, 2000.

[17] Bagging, boosting and stacking in machine learning [Internet], <https://stats.stackexchange.com/questions/18891/bagging-boosting-and-stacking-in-machine-learning>



김 상 훈

<https://orcid.org/0000-0001-8351-459X>

e-mail : kshsang85@naver.com

2012년 숭실대학교 산업·정보시스템공학과 (학사)

2015년 숭실대학교 산업·정보시스템공학과 (공학석사)

2014년~현재 (재)한국화학융합시험연구원(기획조정)

선임연구원

관심분야 : 데이터마이닝, 빅데이터분석, 사업기획·분석



정 병 희

<https://orcid.org/0000-0002-8233-3479>

e-mail : bhchung@ssu.ac.kr

1977년 서울대학교 산업공학과(학사)

1779년 서울대학교 산업공학과(공학석사)

1986년 서울대학교 산업공학과(공학박사)

1980년~현 재 숭실대학교

산업·정보시스템공학과 교수

관심분야: 공급사슬관리, 데이터 마이닝



이 건 호

<https://orcid.org/0000-0001-6943-7911>

e-mail : ghlee@ssu.ac.kr

1996년 U. of Iowa, Dept of Industrial

Eng., Ph.D

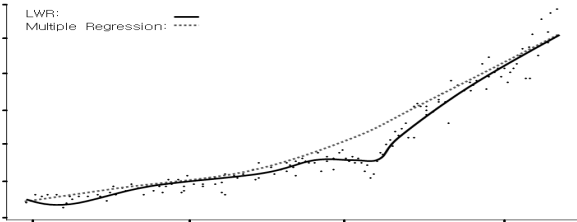
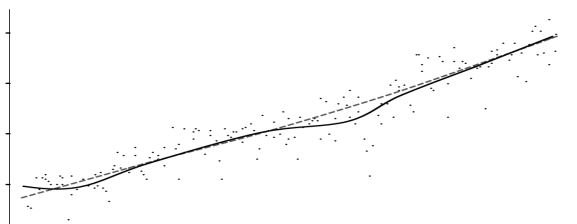
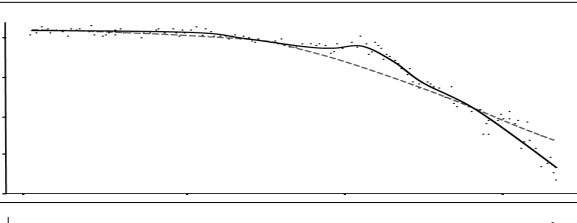
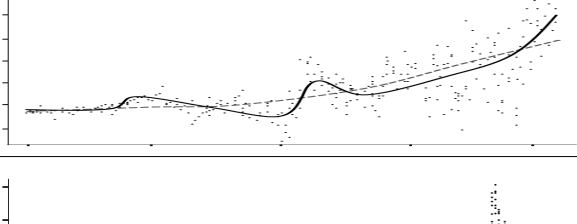
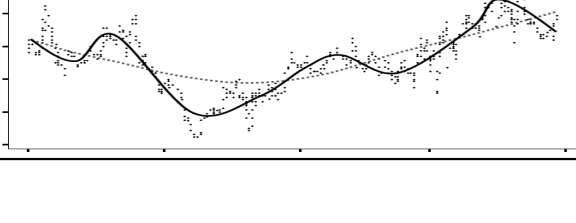
1997년~현 재 숭실대학교

산업·정보시스템공학과 교수

관심분야: 데이터마이닝, 기계학습, 지식기반시스템

Appendix

Table 2. Multiple Regression Patterns and Connected Patterns in This Study

Problem	No. of Data	Comparison of multiple regression pattern LWR patterns connected	Fitness value		CPU time(sec)
			Multiple regression	This study	
1	168		2.15	1.54	30.76
2	260		0.32	0.27	32.83
3	336		0.92	0.54	29.93
4	418		1.31	0.89	83.89
5	1982		2.52	0.49	91.12