선형회귀와 국부적인 RBFN에 의한 점진적인 모델의 설계

이명원*, 곽근창**
*조선대학교 제어계측공학과
**조선대학교 제어계측로봇공학과
e-mail: kwak@chosun.ac.kr

Design of Incremental Model by Linear Regression and Local RBFNs

Myung-Won Lee*, Keun-Chang Kwak**
*Dept of Control and Instrumentation Engineering, Chosun University
**Dept of Control, Instrumentation, and Robot Eng., Chosun University

요 약

본 논문은 선형회귀(LR: Linear Regression)와 국부적인 방사기저함수 네트워크(RBFN: Radial Basis Function Networks)를 결합한 점진적인 모델(incremental model)의 설계와 관련되어진다. 전형적인 RBFN에 의한 모델링과는 달리, 제안된 방법의 근본적인 원리는 두 단계에 의해 고려되어진다. 첫째, 전체 모델의 설계과정에서 전역적인 모델로써 선형회귀에 의해 데이터의 선형부분을 구축한다. 다음으로, 모델링 오차는 오차가 존재하는 국부적인 공간에서 RBFN에 의해 보상되어진다. 여기서, 오차의 분포로부터 RBFN을 설계하기 위해 컨텍스트 기반 퍼지 클러스터링(CFC: Context-based Fuzzy Clustering)를 통해 정보입자의 형태로 구축되어진다. 실험은 자동차 mpg 연료소비량 예측과 부동산 가격예측문제를 통해 제안된 방법의 우수성을 증명한다.

1. 서론

그래뉼러 컴퓨팅은 대량의 데이터, 정보와 지식을 가진 복잡한 응용문제에 대해서, 효과적으로 계산적인 모델을 구축하기 위해 클래스, 클러스터, 부분집합, 그룹, 구간과 같은 정보입자를 효과적으로 이용하는 것에 대한 일반적인 계산 이론방법이다[1]. 이것은 집합이론, 퍼지집합, 러프집합의 환경에서 형성되고 이미 존재하고 잘 정립된 정보입자의 개념으로부터 도출할 수 있는 직접적인 이점을 가지고 있다. Pedrycz[2]는 그래뉼러 컴퓨팅의 개념적이면서 계산적인 플랫폼을 형성하기 위해 퍼지 클러스터링의 근본적인 아이디어를 직접적으로 이용하는 언어적인 모델(LM: Linguistic Model)을 설계하고, 이를 통해 점진적인 모델로 발전시켰다.

이 모델은 입력과 출력공간에서 형성되는 퍼지 집합들 사이에서 관련성을 기술하는 것이며, 이러한 관계를 형성하고 있는 컨텍스트는 시스템 개발자에 의해 정립되어진다.

본 논문에서는 선형회귀(LR: Linear Regression)와 국부적인 방사기저함수 네트워크(RBFN: Radial Basis Function Network)를 결합한 새로운 형태의 점진적인 모델을 설계한다. 여기서 정보입자들은 컨텍스트 기반 퍼지 클러스터링(CFC: Context-based Fuzzy Clustering)[3]을 이용함으로써 구축되어지며, 이러한 방법은 인간 중심형 컴퓨팅 특성을 강조하고 있다. 본 논문에서 제안된 점진적인 모델의 설계원리는 전역적인 모델로써 선형회귀를 이용하여 선형부분을 모델링하고, 선형회귀로부터 얻어진 비선형성을 갖는 모델 오차는 국부적인 방사기저함수 네트워크에 의해 구축되어진다.

제안된 방법의 성능을 평가하기 위해 자동차 연료소비량 예측과 부동산 가격 예측문제를 이용하여 기존의 방법인 LR, RBFN[4], LM[5]과 비교되어진다.

2. 정보입자에 의한 RBFN

본 절에서는 Pedrycz[2]에 의해 소개된 언어적인 모델을 이용하여 그래뉼러 네트워크의 개념을 이끈다. 언어적인 모델은 퍼지 클러스터링의 근본적인 아이디어를 직접적으로 이용하는 퍼지 모델링의 범주에 속한다. 이러한 클러스터링은 정보입자의 개념을 둔 컨텍스트 기반 퍼지 클러스터링에 의해 수행되어진다. 간략히 컨텍스트 기반 퍼지 클러스터링에 대해서 설명하면 다음과 같다. 이 클러스터링의 소속행렬 U는 다음 식과 같이 계산되어진다[3].

$$u_{ik} = \frac{f_k}{\sum_{j=1}^{c}\left(\frac{\| x_k - c_i \|}{\| x_k - c_j \|}\right)^{\frac{2}{m-1}}} \quad (1)$$

여기서 m은 퍼지화 계수이며, f_k의 값은 0과 1사이의 소속도 값을 나타낸다. $f_k = T(d_k)$는 출력공간에서 생성된 임의의 퍼지 집합에서 k번째 데이터의 포함정도를 표현한다. 출력공간의 퍼지 집합은 $T: D \rightarrow [0,1]$로 정의되며, D는

출력변수의 전체집합이다. 여기서 이러한 특성에 의해 소속행렬의 요구조건을 수정하면 다음과 같다.

$$U(f)=\left\{\begin{array}{l} u_{ik}\in[0,1]|\sum_{i=1}^{c}u_{ik}=f_k\,\forall k \\ \text{and } 0<\sum_{k=1}^{N}u_{ik}<N \end{array}\right\} \quad (2)$$

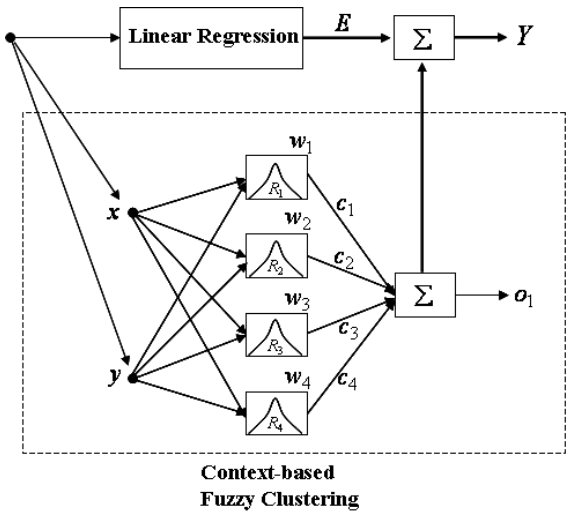
그림 1의 점선부분은 컨텍스트기반 퍼지 클러스터링을 통해 얻어진 RBFN을 보여주고 있다. 전형적으로 R_i는 가우시안 함수이며, 이들의 중심은 클러스터링을 통해 얻어진다. i번째 은닉유닛에 의해 계산되어진 활성함수 w_i는 입력벡터가 그 유닛의 중심에 있을 때 최대값을 얻는다. RBFN의 출력은 다음과 같이 R_i와 관련된 출력 값의 가중치된 합으로써 얻어진다.

$$d(x)=\sum_{i=1}^{H}c_i w_i=\sum_{i=1}^{H}c_i R_i(x) \quad (3)$$

여기서 c_i는 최소자승법에 의해 최종적으로 계산된다.

3. 점진적인 모델의 설계

그림 1은 점진적인 RBFN의 구축에 있어서 전체적인 설계 흐름도를 보여주고 있다. 그림에서 보는 바와 같이 두 단계에 의해 이루어지고 있으며, 선형회귀를 이용하여 전역적인 모델을 구축한 후, 모델의 회귀부분에 의해 얻어진 시스템 모델 오차를 줄이기 위해 국부적인 비선형성을 표현하는 방사기저함수를 통해 국소적인 모델을 구축한다. 컨텍스트기반 퍼지 클러스터링을 이용한 방사기저함수 네트워크를 구축하는 방법은 Pedrycz의 논문을 참고하기 바란다[4].



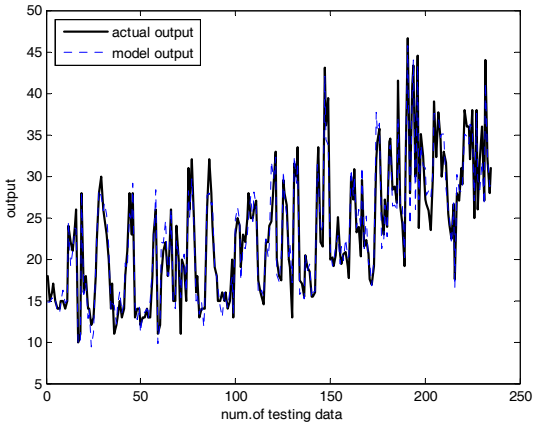
<그림 1> 점진적인 모델의 전체적인 설계 흐름도

4. 실험 및 결과

본 논문에서는 제안된 방법의 성능평가를 위해 자동차 mpg 연료소비량 예측문제와 부동산 예측문제를 다룬다.

4.1 자동차 연료소비량 예측

대표적인 비선형 회귀문제인 자동차 연료 소비량 예측문제를 다룬다. 이 예제에서 6개의 입력이 사용되며 이는 실린더 수, 배기량, 마력, 무게, 가속력, 모델년도로 구성되어진다. 이들 입력에 의해 예측되고자 하는 출력변수는 자동차 연료소비량인 MPG(miles per gallon)이다. 데이터 집합은 392개의 서로 다른 자동차로부터 얻어졌다. 학습데이터와 검증데이터는 0과 1사이로 정규화 되어졌으며, 각각 60%-40%로 임의로 나누어 10번씩 반복적으로 실험하였다. 여기서 학습 데이터는 점진적인 모델을 구축하기 위해 사용되어지며, 검증 데이터는 구축된 모델이 타당한지 살펴보는 모델검증과 관련되어진다. 이렇게 함으로써 결과적인 모델이 학습데이터에 편향되지 않고 새로운 데이터에 대해 좋은 일반화 능력을 가질 수 있다. 또한, 퍼지화 계수(m=2.0), 컨텍스트의 수($2\le p\le 6$), 클러스터의 수($2\le c\le 6$)를 변화해가면서 가장 좋은 성능의 값(p=c=6)을 얻어냈다. 그림 2는 검증데이터에 대한 일반화 능력을 보여주고 있다. 표 1은 학습데이터와 검증데이터에 대한 예측 성능을 비교하고 있다. 표에서 보는 바와 같이 제안된 점진적인 모델은 기존의 방법과 비교해서 우수한 예측 성능을 보이고 있음을 알 수 있었다.



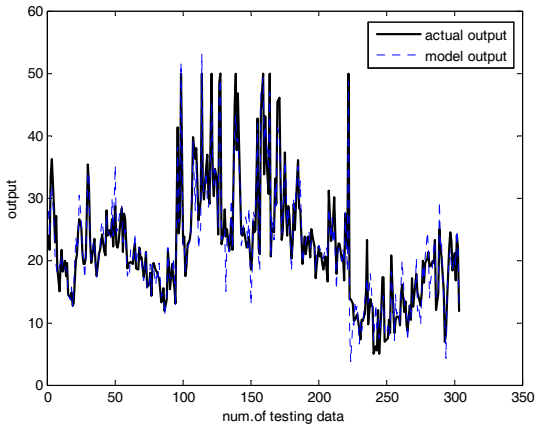
<그림 2> 검증데이터에 대한 일반화 예측능력

<표 1> 학습데이터와 검증데이터에 대한 성능비교

	학습 RMSE	검증 RMSE
LR	3.38	3.47
RBFN	2.33	3.17
LM	2.80	3.32
Incremental model	2.03	3.04

4.2 부동산 가격예측

보스턴의 Housing Data를 사용 하여 부동산 가격 예측문제를 다룬다. 이 예제에서는 13개의 입력 변수 중 본 논문에서는 12개의 입력을 사용한다. 12개의 입력 변수로는 CRIM:각 도시에서의 1인당 범죄 비율, ZN:25,000 sq.ft 이상의 거주 지역 비율, INDUS:도시당 비 소매상업 비율, NOX:일산화질소의 집중도, RM:주거당 평균 방의 수, AGE:1940년 이전에 지어진 집의 비율, DIS:다섯개의 보스턴 고용 중심지까지의 가중 거리, RAD:고속도로까지의 접근 지표, TAX:만 달러당 재산세, PTRATIO: 학생과 선생님 비율, B: 1000(Bk-0.63)^2, 여기서 Bk는 도시 당 흑인의 비율을 의미함. LSTAT: 저소득층 비율로 구성되어진다. 이들 입력에 의해 예측되고자 하는 출력 변수는 BOSTON 외곽지역의 집의 가격(MEDV)이다. 데이터 집합은 Boston 외곽의 506개 지역을 대상으로 조사하여 얻어졌다. 실험방법은 자동차 연료소비량 예측문제와 동일하다. 그림 3은 10번 실험 중 일반화 능력을 보여주고 있는 예측 예를 보여주고 있다. 또한, 표 2는 학습오차와 검증오차를 RMSE로 보여주고 있다. RBFN과 LM은 컨텍스트 기반 퍼지 클러스터링에 의해 설계되었으며, 점진적인 모델과 같이 은닉층과 퍼지 규칙을 각각 36개(p=c=6)를 선택하였다. 표에서 보는바와 같이 점진적인 모델은 기존의 LR, RBFN, LM에 비해 우수한 성능을 보이고 있음을 증명하고 있다.



<그림 3> 검증데이터에 대한 일반화 예측능력

<표 2> 학습데이터와 검증데이터에 대한 성능비교

	학습 RMSE	검증 RMSE
LR	4.58	5.07
RBFN	3.38	4.49
LM	4.56	5.56
Incremental model	2.65	3.99

5. 결론

본 논문에서는 점진적인 모델을 설계하기 위해 전역적인 모델로써 선형회귀로, 지역적인 모델로써 방사기저함수 네트워크를 사용하였다. 또한 방사기저함수 네트워크는 정보입자의 개념에 의한 컨텍스트 기반 퍼지 클러스터링에 의해 구축되어졌다. 비선형회귀의 대표적 모델인 자동차 mpg 연료예측문제와 보스턴의 부동산가격 예측문제에 적용한 실험결과를 살펴보면, 제안된 방법이 기존의 방법들에 비해 근사화 및 일반화 성능이 우수하며 예측성능이 좋음을 알 수 있었다.

감사의 글

본 연구는 교육과학기술부와 산업기술진흥원의 지역혁신 인력양성사업으로 수행된 연구결과입니다. 또한, 본 연구는 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 20090074465)

참고문헌

[1] W. Pedrycz, A. Skowron, V. Kreinovich, *Handbook of Granular Computing*, Wiley, 2008.

[2] W. Pedrycz and K. C. Kwak, "The development of incremental models", *IEEE Trans. on Fuzzy Systems*, Vol. 15, No. 3, pp. 507-518, 2007.

[3] W. Pedrycz, "Conditional fuzzy c-means", *Pattern Recognition Letters* vol. 17, pp.625-632, 1996.

[4] W. Pedrycz, "Conditional fuzzy clustering in the design of radial basis function neural networks", *IEEE Trans. on Neural Networks,* vol. 9, no. 4, pp. 745-757, 1999.

[5] W. Pedrycz and A.V.Vasilakos, "Linguistic models and linguistic modeling", *IEEE Trans. on Systems, Man, and Cybernetics-Part C*, vol. 29, no. 6, pp. 601-612, 1998.