클러스터링 방법을 이용한 TSK 퍼지추론 시스템의 설계 및 해석

오성권*

Design and Analysis of TSK Fuzzy Inference System using Clustering Method

Sung-Kwun Oh*

요 약 본 논문에서는 주어진 데이터 전처리를 통한 새로운 형태의 TSK기반 퍼지 추론 시스템을 제안한다. 제안된 모델은 주어진 데이터의 효율적인 처리를 위해 클러스터링 기법인 Fuzzy C-Means 클러스터링 방법을 이용하였다. 제안된 새로운 형태의 퍼지추론 시스템의 전반부는 FCM 을 통하여 정규화된 멤버쉽 함수와 클러스터 수를 결정하기 때문에, 멤버쉽함수의 형태 및 개수를 정의할 필요가 없어, 모델의 구조 또한 간단한 형태를 이룬다. 본 논문에서 사용된 후반부는 4가지 형태로-간략추론, 1차선형추론, 2차선형추론, 변형된 2차선형추론-가 있으며, 이는 효율적인 후반부구조를 찾는데 주도적인 역할을 한다. 또한 제안된 모델의 후반부 파라미터 계수는 Weighted Least Squares Estimation(WLSE)을 사용하여 동정하며, Least Squares Estimation(LSE)를 적용한 모델의 성능과 비교한다. 마지막으로, Boston housing 데이터를 사용하여 제안된 모델의 성능을 평가하였다.

Abstract We introduce a new architecture of TSK-based fuzzy inference system. The proposed model used fuzzy c-means clustering method(FCM) for efficient disposal of data. The premise part of fuzzy rules don't assume any membership function such as triangular, gaussian, ellipsoidal because we construct the premise part of fuzzy rules using FCM. As a result, we can reduce to architecture of model. In this paper, we are able to use four types of polynomials as consequence part of fuzzy rules such as simplified, linear, quadratic, modified quadratic. Weighed Least Square Estimator are used to estimates the coefficients of polynomial. The proposed model is evaluated with the use of Boston housing data called Machine Learning dataset.

Key Word : Fuzzy c-means Clustering Method, Particle Swarm Optimization, Weighted Least Square Estimator, Least Square Estimator Fuzzy Inference System,

Ⅰ. 서 론

1965년 '퍼지집합'이론[1]은 Zadeh에 의해 처음 소개되었다. 이를 계기로 비선형적이고 다변수인 시스템을 대상으로 한 퍼지 모델링 기법은 이미 잘 알려져 있으며, 이들은 퍼지 추론 시스템의 기초하고 있다.[2] 현재는 '퍼지집합'을 확장한 퍼지 뉴럴 네트워크[3], RBF 뉴럴 네트워크[4] 그리고 퍼지 다항식 뉴럴 네트워크[5] 등 다양한 구조가 시스템 모델링분야에서 연구되고 있다. 그 중에서 퍼지모델의 구조 및 성능개선에 관한 연구가 활발히 진행되고 있다. 하지만 많은 연구에도 불구하고 구조 및 파라미터 동정과 더불어 많은 입력을 사용하였을 경우 발생하는 지나친 퍼지규칙 수의 증가는 여전히 해결해야 할 연구과제로 남아있다. 본 논문에서는 FCM기반

* Corresponding Author: Electronic Engineering Professor of Suwon University (wdkim,ohsk,hkkim@suwon.ac.kr)
Received : August 11, 2014 Revised : August 25. 2014 Accepted : September 11. 2014

퍼지추론 시스템을 제안한다. 멤버쉽 함수를 FCM을 통해 구한다. 따라서, 퍼지규칙은 FCM의 클러스터 수가 되며, 이는 입력이 많아져도 퍼지규칙 수가 증가하지 않고 오로지 클러스터 수만이 퍼지규칙에 영향을 미치게 된다.

II. 본 론

2.1 제안된 모델의 구조

제안된 FCM기반 퍼지추론 시스템의 구조는 그림1과 같다. 일반적인 퍼지추론 시스템은 입력변수의 수를 결정하면 각 입력변수마다 멤버쉽 함수의 수를 정해주어야 하며, 입력변수의 수와 멤버쉽함수의 수를 고려하여 입력공간을 분할하게 된다. 그러나 본 논문에서는 FCM알고리즘을 이용하여 입력변수마다 멤버쉽 함수의 수를 정해주지 않고, FCM의 클러스터 수만으로 입력공간을 분할하는 구조로 모델을 구축하였다.

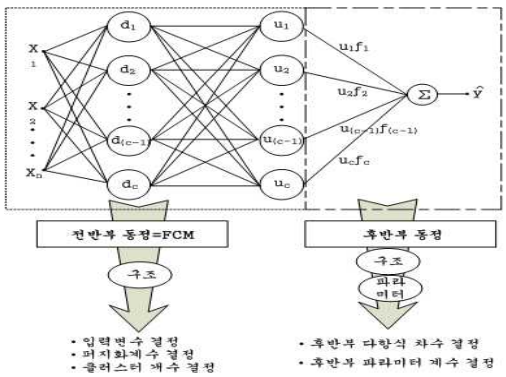

그림 1. FCM기반 퍼지추론 시스템의 구조

2.2 전반부 동정

전반부 동정은 사용하고자 하는 데이터의 입력변수 개수를 선택하고, FCM 알고리즘에 필요한 퍼지화 계수와 클러스터 수를 결정한다.

FCM 알고리즘에서 아래 식(1) 같은 목적함수가 주어진다.

$$u_{ik} = \frac{(1/\|\mathbf{x_k}-\mathbf{v_i}\|^2)^{1/m-1}}{\sum_{j=1}^{c}(1/\|\mathbf{x_k}-\mathbf{v_i}\|^2)^{1/m-1}} = \frac{1}{\sum_{j=1}^{c}(\frac{d_{ik}}{d_{jk}})^{2/m-1}} \quad (1)$$

d_{ik}는 식(2)와 같이 유클리드 거리 법으로 구하지만 본 논문에서는 Manhattan 거리와 Minkowski 거리를 적용하여 성능을 비교하였다.

$$d_{ik} = d(\mathbf{x_k}-\mathbf{v_i}) = [\sum_{j=1}^{s}(x_{ki}-v_{ij})^2]^{1/2} \quad (2)$$

$$d_{ik} = d(\mathbf{x_k}-\mathbf{v_i}) = [\sum_{j=1}^{s}|x_{ki}-v_{ij}|] \quad (3)$$

$$d_{ik} = d(\mathbf{x_k}-\mathbf{v_i}) = [\sum_{j=1}^{s}(x_{ki}-v_{ij})^p]^{1/p} \quad (4)$$

여기서 u_{ik}는 0과 1사이의 수적인 값으로 i번째 클러스터에 속해져 있는 $\mathbf{x_k}$의 k번째 데이터의 소속정도이며, $\mathbf{v_i}$는 i번째 클러스터 중심벡터이다.

j는($j=1,2,...,s$) 특성 공간상의 변수이며, m은 퍼지화 계수이며 논문에서는 m=2로 설정하였다. p는 ($1 \le p \le \infty$) 범위를 가지고 있으며, p=1이면 Manhattan 거리와 같으며, p=2이면 유클리드 거리와 같음을 알 수 있으며, 제안된 모델에서 p는 입력변수의 수와 동일하게 설정하였다.

2.3 후반부 동정

제안된 모델의 후반부 구조는 식(5)~(8)처럼 4가지 형태로 다양하게 적용하였다.

$$R^j : \text{IF } x_1 \text{ is } A_{1c} \text{ and } \cdots \text{ and } x_k \text{ is } A_{kc} \text{ THEN } y_j = f_j(x_1,\cdots,x_k)$$

간략추론[Simplified] :

$$f_j(x_1,\cdots,x_k) = a_{j0} \quad (5)$$

1차선형추론[Linear] :

$$f_j(x_1,\cdots,x_k) = a_{j0} + \sum_{i=1}^{k} a_{ji}x_i \quad (6)$$

2차선형추론[Quadratic] :

$k=2$:

$$f_j(x_1,\cdots,x_k)=a_{j0}+\sum_{i=1}^{k}a_{ji}x_i+\sum_{i=1}^{k}a_{j(k+i)}x_i^2+a_{(2k+1)}x_1x_2$$

$k\geq 3$:

$$f_j(x_1,\cdots,x_k)=a_{j0}+\sum_{i=1}^{k}a_{ji}x_i+\sum_{i=1}^{k}a_{j(k+i)}x_i^2 + a_{(2k+1)}x_1x_2+\cdots+a_{(k(k+3)/2)}x_{(k-1)}x_k \quad (7)$$

변형된 2차선형추론[Modified Quadratic] :

$k=2$:

$$f_j(x_1,\cdots,x_k)=a_{j0}+\sum_{i=1}^{k}a_{ji}x_i+a_{(2k+1)}x_1x_2$$

$k\geq 3$:

$$f_j(x_1,\cdots,x_k)=a_{j0}+\sum_{i=1}^{k}a_{ji}x_i+a_{(2k+1)}x_1x_2+\cdots + a_{(k(k+3)/2)}x_{(k-1)}x_k \quad (8)$$

여기서 $x=[x_1,x_2,\ldots,x_k]$ k는 입력변수의 수, R^j는 j번째 퍼지규칙($j=1,\cdots,c$), c는 퍼지규칙 수이며, $f_j=(x_1,\ldots,x_k)$는 j번째 규칙에 대한 후반부로써 j번째 퍼지규칙에 대한 로컬모델이다.

모델의 출력은 식(9)처럼 구해진다.

$$\hat{y}=\sum_{j=1}^{c}u_j f_j(x_1,\cdots,x_k) \quad (9)$$

후반부 다항식의 계수는 WLSE를 사용하여 구한다. WLSE는 회귀다항식의 계수를 추정하는 알고리즘이며 LSE와 유사하다. LSE는 오차제곱의 합이 최소가 되도록 계수를 추정하지만, WLSE는 오차제곱에 가중치가 곱해진다는 차이가 있다. LSE는 후반부 다항식들의 계수를 한 번에 구하기 때문에 전역 모델의 학습을 수행하게 되고, 입력변수와 멤버쉽 함수의 수가 많아지면 퍼지규칙 수가 기하급수적으로 늘어나기 때문에 컴퓨터 연산 시간이 오래걸리며, 각 퍼지규칙에 대한 해석력도 사라지는 경향이 있다. 그렇지만 WLSE는 각 규칙의 후반부 다항식의 계수를 퍼지규칙마다 독립적으로 구하며 로컬학습을 수행한다. 또한 분할된 입력공간에 적합한 로컬모델을 형성함으로써 각 로컬영역에 대한 해석력을 향상시킬 수 있는 장점이 있다.

식(10)은 WLSE에서의 성능평가함수 Q를 행렬식으로 표현하였다.

$$Q=\sum_{j=1}^{c}(Y-X_j a_j)^T U_j (Y-X_j a_j) \quad (10)$$

여기서, a_j는 추정하고자 하는 j번째 다항식의 계수, Y는 출력데이터, U_j는 j번째 입력공간에 대한 입력 데이터들의 소속 값 을 의미한다. X_j는 j번째 로컬모델의 계수를 추정하기 위한 입력데이터 행렬을 의미하며 로컬모델이 선형일 경우 다음처럼 정의한다.

$$X_j=\begin{bmatrix}1 & x_{11} & \cdots & x_{1k}\\ 1 & x_{21} & \cdots & x_{2k}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{m1} & \cdots & x_{mk}\end{bmatrix} \quad U_j=\begin{bmatrix}u_{j1} & 0 & \cdots & 0\\ 0 & u_{j2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & u_{jm}\end{bmatrix}$$

여기서, m은 데이터의 수이다.

j번째 규칙에 대한 로컬모델인 다항식의 계수는 식(11)에 의해 구해진다.

$$a_j=(X_j^T U_j X_j)^{-1}X_j^T U_j Y \quad (11)$$

Ⅲ. 시뮬레이션

제안된 모델의 성능 평가를 위해서 시스템 모델링에 널리 사용되는 비선형 데이터인 boston housing을 사용했다. 이 데이터는 보스턴 지역 부동산의 정보와 관련이 있으며, 총 506개의 입출력 데이터 쌍으로, 13입력-1출력으로 이루어져있다.

제안된 모델의 평가를 위해 boston housing 데이터를 학습데이터(60%)와 테스트데이터(40%)로 랜덤하게 나누고, 10번 분류한 데이터를 이용해서 구한 성능지수를 평균과 표준편차로 환산하여 나타낸다. 성능지수는 아래 식(12)에 RMSE를 사용한다.

$$RMSE=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2} \quad (12)$$

표 1. WLSE 일 때의 성능지수

No. of Cluster	Distance	Polynomial type	PI	EPI
20	Euclidean	Simplified	7.634±0.284	7.540±0.432
		Linear	2.285±0.102	3.875±0.439
		Quadratic	0.841±0.054	8.234±4.589
		M.Quadratic	0.963±0.067	6.073±0.825
	Manhattan	Simplified	7.472±0.290	7.444±0.454
		Linear	2.464±0.097	3.769±0.225
		Quadratic	1.017±0.055	6.666±3.228
		M.Quadratic	1.160±0.080	5.145±0.651
	Minkowski	Simplified	7.824±0.237	7.743±0.414
		Linear	2.267±0.113	3.930±0.495
		Quadratic	0.763±0.422	8.303±4.024
		M.Quadratic	0.880±0.054	6.533±0.975
25	Euclidean	Simplified	7.518±0.288	7.465±0.468
		Linear	2.186±0.148	3.935±0.466
		Quadratic	0.779±0.039	7.647±4.766
		M.Quadratic	0.911±0.069	6.503±1.239
	Manhattan	Simplified	7.357±0.295	7.313±0.497
		Linear	2.368±0.082	3.853±0.353
		Quadratic	0.957±0.041	6.604±3.003
		M.Quadratic	1.114±0.069	5.297±0.783
	Minkowski	Simplified	7.693±0.276	7.658±0.495
		Linear	2.171±0.150	4.045±0.527
		Quadratic	0.714±0.040	7.902±4.225
		M.Quadratic	0.851±0.070	6.703±0.977

표 2. LSE 일 때의 성능지수

No. of Cluster	Distance	Polynomial type	PI	EPI
20	Euclidean	Simplified	7.485±0.299	7.458±0.402
		Linear	0.940±0.099	218.58±234.1
		Quadratic	0.009±0.031	82.64±42.09
		M.Quadratic	0.016±0.038	82.64±42.09
	Manhattan	Simplified	7.069±0.284	7.307±0.492
		Linear	0.849±0.138	220.02±191.9
		Quadratic	0.007±0.024	109.58±65.86
		M.Quadratic	3.2e-8±3e-8	171.91±127.2
	Minkowski	Simplified	7.725±0.242	7.732±0.360
		Linear	0.949±0.121	184.79±275.1
		Quadratic	0.011±0.034	61.32±20.94
		M.Quadratic	0.014±0.044	216.45±389.5
25	Euclidean	Simplified	7.325±0.285	7.331±0.503
		Linear	0.490±0.160	266.53±236.4
		Quadratic	0.011±0.035	58.93±24.21
		M.Quadratic	0.012±0.040	142.07±184.2
	Manhattan	Simplified	6.943±0.264	7.100±0.584
		Linear	0.343±0.211	311.55±213.6
		Quadratic	0.021±0.067	89.47±58.98
		M.Quadratic	0.001±0.005	80.608±37.21
	Minkowski	Simplified	7.552±0.248	7.589±0.499
		Linear	0.466±0.149	277.10±307.2
		Quadratic	0.012±0.040	56.61±26.71
		M.Quadratic	0.010±0.033	83.18±43.92

표1과 2를 비교하면 WLSE를 사용하면 PI의 성능은 LSE일 때보다 떨어지지만 EPI의 성능이 LSE보다 우수하면 클러스터가 많아져도 성능이 크게 변화하거나 발산하지 않으며 안정적이다.

표3은 기존의 모델과 제안된 모델의 성능을 비교한 것으로 후반부 다항식 차수가 Linear 일 때가 PI 와 EPI의 성능이 우수한 것을 알 수 있다.

표 3. 기존모델과 성능지수 비교

Model		No. of Cluster	PI	EPI
RBFNN[7]		H=25	6.36±0.24	6.94±0.31
RBFNN with context-free clustering[7]		H=25	5.52±0.25	6.91±0.45
Linguistic modeling[7]				
without optimization		H=25	5.21±0.12	6.14±0.28
One-loop optimization		H=25	4.80±0.52	5.22±0.58
Multi-step optimization		H=25	4.21±0.35	5.32±0.96
Our model	Simplified	H=25	7.634±0.284	7.540±0.432
	Linear		2.285±0.102	3.875±0.439
	Quadratic		0.841±0.054	8.234±4.589
	M.Quadratic		0.963±0.067	6.073±0.825

IV. 결 론

본 논문에서는 FCM기반 퍼지추론 시스템을 제안했다. 전반부는 FCM알고리즘을 사용하여 기존의 퍼지추론 시스템보다 구조를 간소화 하였으며, 퍼지규칙 수는 클러스터 수와 같게 되므로 입력이 많은 데이터도 모델링이 가능하다. 후반부 학습은 WLSE를 사용하여 후반부 다항식 계수를 추정하여 로컬 학습 및 퍼지규칙 수를 다양하게 설정하여 모델링이 가능하도록 하였으며, 로컬 모델의 해석력을 향상 시켰다. 앞으로 최적화 알고리즘과 결합시켜 각 시스템에 최적의 구조 및 성능을 갖는 모델을 만드는 방향으로 연구계획을 세웠다.

감사의 글

본 연구는 경기도의 경기도지역협력연구센터 사업의 일환으로 수행하였음[GRRC 수원 2009-B2, U-city 보안감시 기술협력센터]. 그리고 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임[2009-0074928]. .

Reference

[1] L.A. Zadeh “Fuzzy set," Inf. control 8, pp.338-353, 1965.

[2] Mahmut firat, Mustafa Erkan Turan, Mehmet Ali Yurdesev, "Comparative analysis of fuzzy inference systems for water consumption time series prediction," Journal of Hydrology, Vol.374, No.3-4, pp.235-241, 2009.

[3] Cheng-jian Lin, "An efficient immune-based symbiotic particle swarm optimization learning algorithm for TSK-type neuro-fuzzy networks design," fuzzy ets and Systems, Vol.159, No.21, pp.2890-2909, 2008.

[4] A. Stajano., j. Tagliaferri, W. Pedrycz, "Improving RBF networks performance in regression tasks by means of a supervised fuzzy clustering Automatic structure and parameter," Neurocomputing, Vol.69, pp.1570-1581, 2006.

[5] H.-S. Park, W. Pedrycz, S.-K. Oh, "Evolutionary design of hybrid self-organizing fuzzy polynomial neural networks with the aid of information granulation," ESWA, Vol.33, No.4, pp.830-846, 2007.

[7] W. Pedrycz. K.C. Kwak, "Linguistic models as a frameworks of user-centric system modeling," IEEE, Trans. SMC-A, Vol.36, No.4, pp.727-745, 2006.

저자약력

오 성 권(Sung-Kwun Oh) **정회원**

1981년 연세대학교 전기공학과 졸업.
1983년 동 대학원 전기공학과 졸업(공학석사)
1983~1989년 금성산전연구소(선임연구원)
1993년 연세대 대학원 전기공학과 졸업(공학박사)
1996~1997년 캐나다 Manitoba 대학 전기 및 컴퓨터공학과 Post-Doc
1993~2005년 원광대 전기전자 및 정보공학부 교수
2005년~현재 수원대 전기공학과 교수
2002년~현재 : 대한전기학회, 퍼지및지능시스템학회및제어자동화시스템공학회 편집위원.

<관심분야> 시스템 자동화, 퍼지이론, 신경회로망 응용 및 제어, 컴퓨터 지능 등.