# 퍼지 신경망을 이용한 공간 분류 시스템의 설계 및 구현

안찬민* 박상호 박태수 이주홍
인하대학교 컴퓨터정보공학과
{ahnch1*, parksangho, taseu}@datamining.inha.ac.kr
juhong@inha.ac.kr

# Design and Implementation of Spatial Classification System using Fuzzy-Neural Network

Chan-Min Ahn*, Sang-Ho Park, Tae-Su Park, and Ju-Hong Lee
Department of Computer Science & Information Engineering
Inha University, Incheon, Korea

요 약

기존 공간 분류 시스템은 애매모호한 데이터나 불완전한 데이터, 결손 데이터의 처리에는 취약하다는 단점을 가지고 있다. 수치 형태의 애매모호성을 효과적으로 처리하기 위해 신경망을 이용할 수 있다. 그러나, 신경망을 이용한 공간 데이터 분류 방법은 불완전한 데이터나 결손 데이터들을 무시하지 않고 처리할 수 있으나, 다양한 수치형태를 가지는 공간 데이터들로 인해 네트워크 구조의 복잡도가 증가하고 학습 성능이 저하된다는 문제점을 야기한다. 본 논문에서는 이러한 문제점을 해결하기 위해서 퍼지 신경망을 적용한 새로운 공간 분류시스템을 제안하고 구현하였다. 실험 결과 기존의 방법에 비해 좋은 성능을 보임을 확인하였다.

## 1. 서 론

최근의 지리 정보 시스템(GIS)은 단순히 지리 정보를 저장하고 검색하는데 이용할 뿐 아니라 고객 관리, 날씨 및 주변 환경 예측, 교통량의 관리와 같은 다양한 분야에서 지리 정보 종합 관리 시스템으로 이용되고 있다. 따라서 많은 양의 정보를 효율적으로 관리 및 이용하기 위한 연구가 활발히 진행되고 있다.

공간 데이터마이닝 방법 중 공간 분류 방법(Spatial Classification Method)은 공간 데이터베이스에 저장된 데이터의 속성을 분석하여 주어진 클래스 집합으로 분류하는 기법이다[1]. 기존 공간 분류 방법은 영상 데이터의 분류를 위한 방법이 대부분이다 [2,3]. 지리 정보 시스템에서 사용되는 대표적인 공간 분류 방법으로 의사 결정 트리 방법(Decision Tree Method) [1,4], 베이지안 분류자(Bayesian classifier)와 의사 결정 트리를 병용한 방법 [3] 등이 있다. 의사 결정 트리를 기반으로 확장된 공간 분류 방법은 트리에 정의된 분류 기준을 이용하기 때문에 트리의 분류 기준에 맞지 않는 애매모호한 데이터와 트리에 정의되어 있지 않은 결손 데이터를 입력 받을 경우 정확히 분류하지 못하는 단점을 가진다.

본 논문에서는 퍼지 신경망을 이용하여 GIS의 지리 정보 데이터를 분류하는 시스템을 제안하고 구현하였다. 그리고 이를 활용하기 위해 공간 데이터 마이닝 질의 언어인 SIMQL[5]을 확장하였다.

## 2. 관련 연구

대표적인 지리 정보를 위한 공간 분류 시스템은 다음과 같다.

K. Koperski[6]는 의사 결정 트리 방법을 이용하여 데이터베이스에서 비공간 속성 정보를 이용하여 분류된 객체들을 구하고, 분류된 객체들과 다른 객체들 사이의 공간 정보를 나타내는 속성, 함수들을 정의하였다. 그리고 RELIEF 알고리즘을 이용해서 최근접 이웃 탐색(Nearest Neighbor Search)를 수행한다. 이후 임계값(Threshold)을 만족하는 값만을 남긴다. D. Li[3]는 베이지안 분류자와 C5.0 알고리즘을 이용한 분류 방법을 제안하였다. 이 방법은 C5.0을 이용하여 의사 결정 트리와 규칙을 얻는다. C5.0 알고리즘은 ID3에서 발전된 의사 결정 트리 방법으로 대용량의 데이터베이스에서 다중 클래스 분류 문제를 빠르게 수행해주는 방법이다.

그러나 의사 결정 트리를 이용한 공간 분류 방법들은 정의되지 않은 모호하거나 불완전한 공간 데이터를 분류하기 어렵다.

## 3. 공간 분류 시스템

본 논문에서 제안한 공간 분류 시스템은 그림 1과 같이 학습단계와 분류단계로 구성된다. 그림1에서 Parser는 사용자 질의를 분석하여 학습과 분류에 사용될 데이터를 얻는다. Evaluator는 사용자 질의어의 선택 사항들과 수

집된 데이터를 이용하여 학습과 분류 작업을 각각 수행한다.

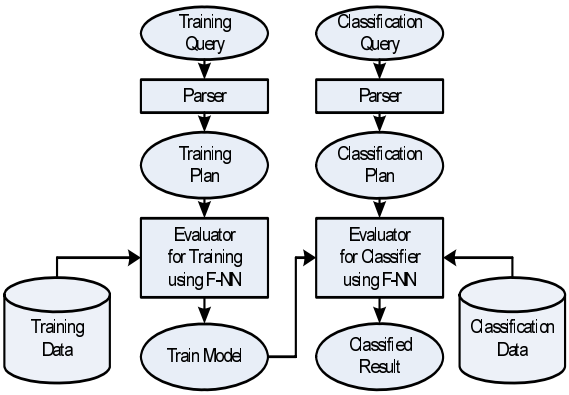


그림 1 공간 분류 시스템의 구조

그림 1의 좌측부분은 학습 단계, 우측 부분은 분류 단계를 각각 보여준다. 학습 단계에서는 퍼지 신경망을 이용하여 학습된 모델을 생성하고, 분류 단계에서는 학습 단계에서 얻은 분류 모델을 이용하여 실질적인 분류 작업을 수행하게 된다.

## 4. 공간 분류 질의 언어

본 논문은 그림 1과 같은 공간 분류 시스템의 학습과 분류단계를 위한 질의의 규격을 확장 BNF형식으로 정의하였다. 본 논문에서는 공간 데이터 마이닝 질의 언어인 SIMQL[5]을 확장하였다. 공간 분류 시스템을 위한 SIMQL은 다음과 같은 특징을 가진다.

- 학습단계와 분류단계는 각각 다른 구문을 사용한다.
- 학습단계에서 학습된 모델을 저장할 수 있다. 반면, 분류단계에서는 저장된 학습된 모델 중 분류에 적용하고자 하는 모델을 선택할 수 있다.
- 학습단계에서 구문을 이용하여 학습데이터의 각 속성 값들이 몇 개의 linguistic term으로 표현되는지를 지정할 수 있다.
- 학습단계에서 WITH SAMPLING OPTION sampling_option 구문을 이용하여 Random 혹은 Hold-Out 혹은 k-fold같은 다양한 샘플링 방법에 의하여 학습 데이터들을 추출할 수 있다.

## 5. 공간 분류 시스템을 위한 사용된 퍼지 신경망

퍼지 신경망 방법은 정보의 모호성을 수학적 이론으로 해석하는 퍼지 이론과 학습을 통한 분류 방법인 신경망 방법을 통합한 방법이다. 퍼지 신경망 모델은 공간 분류 방법뿐 아니라 전반적인 분류 방법에 적용 가능한 모델이다. 대표적인 퍼지 신경망 모델로 NEFCLASS[7], ANFIS[8,9], GARIC[10] 등이 있다

본 논문에서 제안한 공간 분류 시스템의 분류자는 3-레이어 구조의 퍼지 신경망을 이용하여 학습과 테스트를 수행한다. 학습단계에서 퍼지 신경망은 각 레이어에서 퍼지화(Fuzzification), 퍼지규칙(Fuzzy Rule), 비 퍼지화(Defuzzification) 과정을 수행한다. 그림 2는 본 논문에서 제안한 공간 분류 시스템을 위한 3-레이어의 퍼지 신경망의 구조를 보여준다.

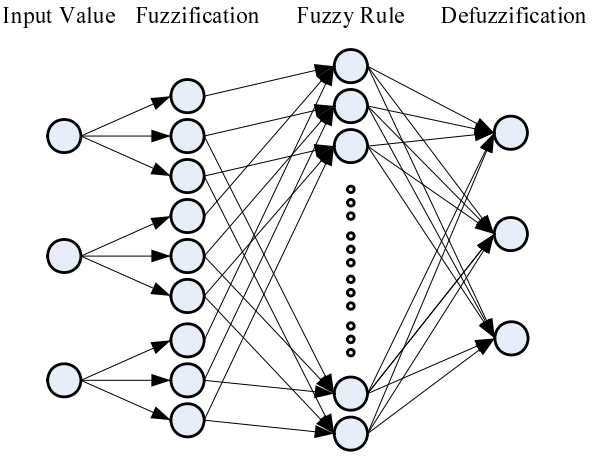


그림 2 퍼지 신경망의 구조

퍼지 신경망은 퍼지 이론를 이용하여, 수치적 형태의 공간 정보를 linguistic Term으로 변환할 수 있으며, 애매모호한 값을 효과적으로 처리하여 학습 능력을 향상시킬 수 있다. 또한, 네트워크 학습에 의해 얻어진 모델을 linguistic term들에 의한 IF-THEN 형태의 룰로 표현할 수 있다. 다음은 *xp*와 *xq*를 원소로 가지는 2차원 입력벡터 *X*가 멤버쉽 함수 F1과 F2에 적용되어, C1 혹은 C2 중 하나의 클래스에 분류될 경우의 학습된 분류 룰을 보여준다.

Rule_1 :
**If** *xp* is high in F1 **AND** *xq* is very low in F2
**Then** *X*=(*xp*,*xq*) belong to C1.
Rule_2 :
**If** *xp* is very high in F1 **AND** *xq* is low in F2
**Then** *X*=(*xp*,*xq*) belong to C2

위와 같은 형태의 룰은 분류 룰을 학습할 때에 퍼지 신경망의 다양한 파라미터들의 학습에 의해 갱신됨으로써 만들어진다. 다음은 그림 2의 퍼지 신경망 구조상의 각 레이어별 구조 학습과 파라미터 학습을 방법에 대한 설명이다.

**첫 번째 레이어(퍼지화 레이어)** : 첫번째 레이어에서는 사용자 질의에 의해 샘플링된 각 데이터들의 속성값들을 입력으로 사용하여 퍼지 멤버쉽 함수에 의한 소속정도를 출력한다. 이 때, 각 속성값은 그 속성의 linguistic term의 개수와 동일한 개수의 퍼지 멤버쉽 함수(Fuzzy membership function) 에 적용된다. 본 논문에서 사용된 퍼지 신경망은 식(1)과 같은 세 개의 파리미터 $a_{ji}$, $b_{ji}$,와 $c_{ji}$를 가지는 일반화된 Bell Shaped 멤버쉽 함수를 이용한다[9]. 여기에서, $a$는 멤버쉽 함수의 퍼짐 정도를 의미하고, $b$는 bell의 형태, 그리고 $c$는 중심값을 의미한다. $b$는 일반적으로 양수 값을 가진다. 다음 식(1)은 j번째 룰을 위한 i번째 입력값에 대한 일반화된 Bell-shaped 함수를 나타낸다.

$$R_{ij}(x_i) = \left(1 + \left(\left(x_i - c_{ij}\right)/a_{ij}\right)^{2b_{ij}}\right)^{-1} \quad (1)$$

**두 번째 레이어(퍼지 규칙 레이어)** : 두번째 레이어에서는 $R^N$개의 규칙(Rule)들에 대한 Firing strength를 출력한다. 본 논문은 식(2)를 적용하여 각 룰의 호환성(compatibility)을 나타내는 Firing Strength를 계산한다. $L_i$는 i번째 규칙 노드를 나타낸다.

$$L_k = \prod_{i=1}^{N} R_{ij}(x_i) \quad , \quad j = \varphi(k,i) \quad (2)$$

여기서 $j$는 $k$번째 룰 노드에 대한 $i$번째 입력 속성변수의 퍼지 linguistic term을 결정하는 인덱스이다.

**세 번째 레이어(비 퍼지화 레이어)**: 세 번째 레이어는 정해진 네트워크 구조상에서 입력된 데이터들이 어떤 클래스에 분류되는가를 분류 결정 단계와 파라미터들의 갱신여부를 결정하는 갱신 단계로 구분된다. 분류 결정 단계에는 다음과 같은 두 단계로 구성된다.

첫째, 각 룰의 Firing Strength와 가중치의 선형 조합을 계산하여 규칙과 클래스간의 적합한 정도(matching degree)를 식 3과 같이 계산한다. $W_{kl}$는 퍼지 rule 레이어의 $k$번째 노드와 비퍼지화 레이어의 $l$번째 노드사이의 가중치를 의미한다.

$$T_l = \sum_{k=1}^{m} w_{kl} L_k \quad (3)$$

둘째, 식(3)에 의해 계산된 적합한 정도를 다시 비퍼지화 레이어의 활성 함수(activation function) 에 적용하여 class의 소속정도를 최종적으로 출력한다. 본 논문에서는 시그모이드 함수(sigmoidal function)를 비퍼지화 레이어의 활성 함수로 이용하였다.

$$O_l = \left[1 + \exp(-T_l)\right]^{-1} \quad (4)$$

갱신 단계는 다음과 같은 단계들에 의하여, 파라이터들의 갱신여부을 결정한다. 먼저, 식(4) 의 결과값과 주어진 입력값들의 레이블값과의 차이를 이용한 식(5)의 최소자승오차(Least Square Error) 함수를 이용하여 오차를 계산한다.

$$E = E(a_{ij}, b_{ij}, c_{ij}, w_{kl}) = \frac{1}{2}\sum_{l=1}^{L}\left(O_l^d - O_l\right)^2 \quad (5)$$

식(5)에 의하여 계산된 에러값이 주어진 임계치보다 큰 경우는 최적화 방법(steepest descent method)을 이용한 식(6), (7), (8), (9)에 의해 $a$, $b$, $c$, $w$의 파라미터들이 갱신된다.

$$a_{ij}(t+1) = a_{ij}(t) + \eta_a(\partial E / \partial a_{ij}) \quad (6)$$

$$b_{ij}(t+1) = b_{ij}(t) + \eta_b(\partial E / \partial b_{ij}) \quad (7)$$

$$c_{ij}(t+1) = c_{ij}(t) + \eta_c(\partial E / \partial c_{ij}) \quad (8)$$

$$w_{li}(t+1) = w_{li}(t) + \eta_l(\partial E / \partial w_{li}) \quad (9)$$

그러므로, 위와 같은 학습에 의하여 갱신된 파라미터들과 가중치 집합은 주어진 학습 데이터에 대한 하나의 모델을 생성하게 된다.

## 6. 실험

본 절에서는 논문에서 제안된 퍼지 신경망 분류 시스템에 대한 성능을 다양한 실험을 통하여 검증한다. 실험 데이터로서 집의 평수, 인구, 소득 등의 비공간 속성과 역까지의 거리, 도로까지의 거리 등의 공간 속성을 사용하였다. 최대 반복되는 학습의 횟수는 1000 에폭(epoch)으로 제한하였으며, 수렴 오차 한계는 0.01로 설정하였다. 예를 들어 어떤 도시를 집값에 따라 분류한다고 가정하면, 학습 질의를 다음과 같이 표현할 수 있다.

```
MINE SPATIAL CLASSIFICATION TRAINING AS house_cost
SAVING TO MODEL NAME house_cost_training
WITH CLASS LABEL expensive not_expensive cheap
FUZZIFIER LINGUSTIC_TERM 3 EACH
FOR distance_to_station, distance_to_market, distance_to_river, distance_to_park, area
FROM census_city
```

이 질의를 파싱(Parsing) 하여 학습에 필요한 데이터

와 파라미터들을 설정하여 퍼지 신경망으로 학습 시킨다. 수렴할 때까지 학습을 반복하여 최적의 멤버쉽 함수와 가중치 값을 찾을 수 있다.

퍼지 신경망을 이용하여 학습된 네트워크를 생성하면, 실험 데이터를 이용하여 분류 정확성을 측정한다. 학습을 통해 얻어진 모델을 이용하여 실험데이터의 분류를 수행하는 질의는 다음과 같다.

```
MINE SPATIAL CLASSIFICATION AS test
USING MODEL NAME house_cost_training
FROM census_city
```

표 1은 의사 결정 트리와 퍼지 신경망에 대한 실험 결과를 나타낸 것이다. 본 실험에서는 2종류의 의사 결정 트리를 사용하였다. 첫 번째는 가지치기(Pruning)를 사용한 의사 결정 트리의 결과이고 다른 하나는 가지치기를 사용하지 않은 의사 결정 트리의 결과이다.

표 1 의사결정트리와 퍼지신경망의 비교 실험 결과

| | 가지치기 안 된 의사 결정 트리 | 가지치기 된 의사결정트리 | 퍼지 신경망 |
|---|---|---|---|
| 정확도 | 65% | 75% | 90% |

성능을 비교하기 위해서 분류정확도를 평가하였다. 그 결과, 퍼지 신경망이 두 종류의 의사 결정 트리에 비해 좋은 정확도를 보였다. 즉, 퍼지 신경망이 적절한 클래스로 보다 정확하게 데이터를 분류할 수 있다.

## 7. 결론

본 논문에서는 지리 정보 시스템에 사용되는 퍼지 신경망을 이용한 공간 분류 시스템을 설계 및 구현하였다. 우리는 본 논문에서 제안한 시스템이 지리 정보 시스템에서 발생할 수 있는 애매모호한 값이나 불완전한 데이터, 결손 데이터를 처리할 수 있음을 보이기 위해 기존의 의사 결정 트리 방법과 비교 실험한 결과를 보였다. 그 결과 본 논문에서 제안한 방법이 의사 결정 트리 방법에 비해 높은 정확도를 갖는 분류 결과를 제공한 것을 확인할 수 있다. 그리고 제안한 분류 시스템을 위한 질의 언어를 정의하였다. 본 논문에서는 정의한 학습 질의 언어와 분류 질의 언어로 구성된 SIMQL 질의 언어를 확장하였다. SIMQL 질의 언어는 퍼지 신경망의 학습에 필요한 데이터뿐 아니라 추출 방법까지 제시할 수 있고 학습된 네트워크를 새롭게 분류가 필요한 데이터에 적용할 수 있도록 구성되었다.

## 부록

본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT 연구센터 육성.지원사업의 연구결과로 수행되었음.

## 참고문헌

[1] Krzysztof Koperski, "A Progressive Refinement Approach to Spatial Data Mining", SIMON FRASER UNIVERSITY, 1999

[2] C.Z. Van de Beek, d R. Uijlenhoet and I. Holleman, "Spatial classification of precipitation from operational radar data", *32nd Conference on Radar Meteorology*, 2005

[3] Li, D, K. Di, D. Li, "Land use classification of remote sensing image with GIS data based on spatial data mining techniques", *19th ISPRS Congress*, Amsterdam, July 16-23,2000.

[4] Q. Ding, Q. Ding, and W. Perrizo, "Decision Tree Classification of Spatial Data Streams Using Peano Count Trees", *Proceedings of the ACM 124 Symposium on Applied Computing*, Madrid, Spain, pp. 413-417, March 2002,

[5] 박선, 박상호, 안찬민, 이윤석, 이주홍: "SIMS를 위한 공간 데이터 마이닝 질의 언어.", 한국정보과학회 춘계학술발표대회, Vol.31,No.1 70-72, 2003

[6] Krzysztof Koperski, Jiawei Han, and Nebojsa Stefanovic, "An Efficient Two-Step Method for Classification of Spatial Data", *Proc. Int. Symp. on Spatial Data Handling*, Vancouver, Canada,1996

[7] Detlef Nauck and Rudolf Kruse, "NEFCLASS-A Neuro-Fuzzy Approach For The Classification of Data", *ACM Symposium on Applied Computing*, 1995

[8] J. S. R. Jang., "ANFIS: Adaptive-network-based fuzzy inference system", IEEE Trans. Syst., Man, Cybern., vol. 23, pp. 665-685, June 1993.

[9] J. S. R. Jang., C.T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing*, Prantice Hall, 1997.

[10] H. R. Berenji and P. Khedkar, "Learning and tuning fuzzy logic controllers through reinforcements", *IEEE Trans. Neural Networks*, vol. 3. pp. 724-740, October 1992.