# FCM 클러스터링 기반 지도 학습 알고리즘을 이용한 당뇨병 예측 분석

박태언 · 김광백*

신라대학교

# Diabetes Predictive Analytics using FCM Clustering based Supervised Learning Algorithm

Tae-eun Park · Kwang-baek Kim

Silla University

E-mail : ase6733@naver.com / gbkim@silla.ac.kr

## 요 약

본 논문에서는 데이터를 정량화하여 특징을 분류하기 위한 방법으로 퍼지 클러스터링 기반 지도 학습 방법을 제안한다. 제안된 방법은 FCM 클러스터링을 기법을 적용하여 군집화를 수행한다. 그리고 군집화 된 데이터들 중에서는 정확히 분류되지 않은 데이터가 존재하므로 분류되지 않은 데이터에 대해 지도 학습 방법을 적용한다. 본 논문에서는 당뇨병의 유무를 타겟 데이터로 설정하고 나머지 8개의 속성의 데이터를 FCM 기반 지도 학습 방법을 적용하여 당뇨병의 유무를 예측한다. 당뇨병 예측에 대한 성능을 30회의 K-겹 교차검증 (K-Fold Corss Validation)을 이용하여 평가하였으며, 다층 퍼셉트론의 경우에는 훈련 데이터가 77.88%, 테스트 데이터가 62.78%로 나타났고 제안된 방법의 경우에는 훈련 데이터가 79.96%, 테스트 데이터 74.16%로 나타났다.

키워드

FCM, 단층 퍼셉트론, 다층 퍼셉트론, 당뇨병

## Ⅰ. 서 론

다층 퍼셉트론은 입력 패턴을 가중치와 바이어스 값을 이용하여 연산하는 퍼셉트론을 하나의 노드가 2개 이상의 노드와 레이어로 구성되어 학습하는 지도 학습 방법 중 하나로 비선형 문제를 해결 하지 못했던 단층 퍼셉트론의 문제를 해결하는 방법으로 제안되었다. 그러나 다층 퍼셉트론의 경우에는 노드의 수와 레이어의 수, 학습 구조의 설계 과정에 따라 성능이 달라진다. 그리고 문제에 따른 구조 설계 방법이 명확하게 존재하지 않아 NP (Non-deterministic Polynomial) 문제로 분류된다는 문제점과 Vanishing Gradient, Over Fitting, Local Minimum 등의 단점을 가지고 있다[1].

FCM (Fuzzy C-Means) 클러스터링은 하나의 클러스터에 속해져 있는 각각의 데이터에 클러스터에 대한 소속 정도를 일일이 열거하여 군집화 하는 비지도 학습 기법이다. 그러나 모든 데이터의 소속 정도를 열거함에 있어 잡음에 약하다는 문제점을 가지고 있다.

본 논문에서는 FCM 클러스터링을 적용하여 데이터를 군집화 한다. 군집된 데이터들 중에서는 정확히 분류되지 않은 데이터가 존재하므로 분류되지 않은 데이터에 대해 지도 학습 방법을 적용하여 안정성의 특징을 가진 지도학습 방법을 제안한다.

## Ⅱ. FCM (Fuzzy C-Means)

FCM (Fuzzy C-Means)은 각각의 클러스터에 대한 데이터의 소속정도를 일일이 열거한 클러스터링 알고리즘이다[2]. 따라서 잡음 데이터의 경우에도 클러스터에 대한 소속정도를 가지기 때문에 잡음에 민감하다는 단점이 있다. 그리고 클러스터의 개수, 지수가중치 설정 등의 하이퍼 파라미터에 따라 성능이 달라지는 단점이 있다.

---

* speaker

FCM (Fuzzy C-Means)

$\epsilon = 0.0001$

$m = 2$

**Step 1**. 지수가중치 $m(1 \leq m < \infty)$을 초기화 한다. 그리고 클러스터의 중심 벡터를 임의의 값으로 초기화 한다.

**Step 2**. 클러스터의 중심벡터와 데이터간의 거리를 식 (1)과 같이 계산한다.

$$d_{ik} = \sqrt{(x_k - v_i)^{2/(m-1)}} \quad (1)$$

식 (1)에서 $d_{ik}$는 데이터와 클러스터간의 거리이다. 그리고 $x_k$는 $k$번째 데이터 이며, $v_i$는 $i$번째 클러스터의 중심 벡터이다.

**Step 3**. 식 (1)을 이용하여 데이터와 중심벡터 간의 거리를 구한 후 각 데이터마다 클러스터의 소속도를 식 (2)과 같이 계산한다.

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

식 (2)에서 $u_{ik}$는 현재 데이터의 소속도를 의미하고 $c$는 클러스터의 개수를 의미한다.

**Step 4**. 식 (2)를 이용하여 데이터가 각각의 클러스터에 소속된 정도를 구하고 식 (1)과 (2)를 통하여 구한 거리와 소속도를 바탕으로 중심 벡터를 식 (3)과 같이 계산한다.

$$v_{ij} = \frac{\sum_{k=1}^{n} (u_{ik})^2 x_{kj}}{\sum_{k=1}^{n} (u_{ik})^2} \quad (3)$$

식 (3)에서 $v_{uj}$는 두 개의 좌표 $i, j$로 이루어진 중심 벡터이며 $x_{kj}$ FCM에 적용된 데이터이며 $n$은 데이터의 총 개수 이다.

**Step 5**. 식 (3)을 이용하여 새롭게 구한 클러스터의 중심 벡터와 이전 중심 벡터의 차이가 $\epsilon$보다 크다면 Step 2로 돌아간다.

FCM 클러스터링에 적용된 데이터는 Diabetes(당뇨병) 데이터 셋으로 임신횟수, 포도당, 혈압, 피부두께, 인슐린, BMI, 당뇨 유전점수, 나이로 총 8개의 속성을 가진 768명의 의료 정보이다[3].

## III. 제안된 방법

제안된 방법에서는 FCM 클러스터링을 이용하여 Diabetes 데이터 셋을 분류하여 정량화 한다. FCM 클러스터링의 경우에는 클러스터의 개수를 정적으로 설정한다. 표 1은 클러스터의 개수에 따른 분류 결과를 나타내었다.

표 1. 클러스터 개수 별 FCM 분류 결과

| | C = 2 | | C = 3 | | C = 4 | | C = 5 | |
|---|---|---|---|---|---|---|---|---|
| 당뇨병 유 / 무 | 유 | 무 | 유 | 무 | 유 | 무 | 유 | 무 |
| 클러스터 1 | 178 | 148 | 29 | 289 | 22 | 252 | 86 | 74 |
| 클러스터 2 | 90 | 360 | 128 | 111 | 111 | 98 | 22 | 251 |
| 클러스터 3 | | | 101 | 111 | 78 | 47 | 15 | 53 |
| 클러스터 4 | | | | | 57 | 100 | 46 | 51 |
| 클러스터 5 | | | | | | | 99 | 71 |

표 1에서 클러스터의 개수를 4개로 설정하여 FCM 클러스터링을 적용한 경우에는 1번째 클러스터에서 당뇨병에 걸리지 않은 환자의 데이터가 밀집되어 상대적으로 효과적인 군집화가 되어 있는 것을 확인 할 수 있다. 그러나 나머지 클러스터 개수에서는 밀집도가 비슷하거나 그 차이가 적어 효과적이지 못한 것을 확인 할 수 있다.

따라서 본 논문에서는 FCM 클러스터링의 클러스터 개수를 4개로 설정하였다.

설정된 클러스터의 개수는 단층 퍼셉트론의 노드의 개수와 동일하며, 입력 데이터가 소속된 노드를 대상으로 학습이 진행된다. 제안된 방법의 구조는 그림 3과 같다.

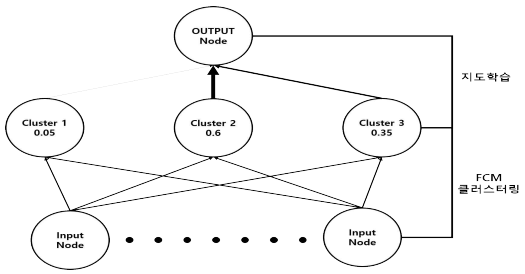


그림 1. 제안된 방법의 구조

그림 1에서 입력된 데이터의 소속 정도에 따라 학습되는 노드가 다른 것을 확인 할 수 있다. FCM 클러스터링의 경우에는 각각의 클러스터에 대한 소속 정도가 열거되기 때문에 1개 이상의 노드에서 입력 데이터에 대한 학습이 진행된다.

## Ⅳ. 실험 및 결과 분석

본 논문에서는 제안된 방법의 성능을 분석하기 위하여 AMD Ryzen 3 3300X 4-Core Processor, 3.79 GHz와 16GB RAM이 장착된 PC에서 Visual Studio 2019 C#으로 제안된 방법을 구현하였다. 제안된 방법의 성능을 검증하기 위해 다층 퍼셉트론을 이용한 당뇨병 예측 방법, 제안된 방법을 이용한 당뇨병을 예측 방법을 각각 실험하여 계층별 K-겹 교차 검증(Stratified K-Fold Cross Validation)을 이용하여 성능을 평가하고 분석하였다. 실험에 사용된 훈련 데이터와 테스트 데이터는 7:3으로 나누어 30회의 계층별 K-겹 교차 검증을 적용하여 평가하였다.

30회의 계층별 K-겹 교차 검증을 이용하여 평가한 다층 퍼셉트론의 평균 성능을 표 2로 나타내었다.

표 2. 다층 퍼셉트론 성능

| 훈련 데이터 | | | |
|---|---|---|---|
| 민감도 | 정확도 | 정밀도 | F1 Score |
| 75.76% | 80.22% | 85.00% | 77.88% |
| **테스트 데이터** | | | |
| 60.88% | 65.23% | 74.96% | 62.78% |

다층 퍼셉트론을 Diabetes 데이터 셋에 적용한 경우에는 훈련데이터에서 민감도가 75.76%, 정확도 80.22%, 정밀도 85.00%, F1 Score가 77.88%로 나타났고 테스트 데이터에서는 민감도가 60.88%, 정확도 65.23%, 정밀도 74.96%, F1 Score가 62.78%로 나타났다.

30회의 계층별 K-겹 교차 검증을 이용하여 평가한 제안된 방법의 평균 성능을 표 3로 나타내었다.

표 3. 제안된 방법 성능

| 훈련 데이터 | | | |
|---|---|---|---|
| 민감도 | 정확도 | 정밀도 | F1 Score |
| 85.05% | 75.49% | 85.10% | 79.96% |
| **테스트 데이터** | | | |
| 78.79% | 70.27% | 80.96% | 74.16% |

제안된 방법을 Diabetes 데이터 셋에 적용한 경우에는 훈련데이터에서 민감도가 85.05%, 정확도 75.49%, 정밀도 85.10%, F1 Score가 79.96%로 나타났고 테스트 데이터에서는 민감도가 78.79%, 정확도 70.27%, 정밀도 80.96%, F1 Score가 74.16%로 나타났다.

## Ⅴ. 결 론

본 논문에서는 FCM 클러스터링을 이용하여 데이터를 분류하고 정량화 하여 지도학습에 안전성을 더하고 Over Fitting의 문제를 최소화 하였다.

제안된 방법의 성능을 평가하기 위해 다층 퍼셉트론과 제안된 방법을 Diabetes 데이터 셋에 적용하여 당뇨병의 예측 성능을 나타내었다. 실험 데이터와 테스트 데이터는 7:3으로 구성하고 30회의 K-겹 교차검증을 이용하여 성능을 평가하였다.

제안된 방법의 경우에는 훈련 데이터의 성능이 다층 퍼셉트론의 성능 보다 2.08% 효과적인 것으로 나타났고 테스트 데이터에서는 다층 퍼셉트론보다 11.38% 효과적인 것으로 나타났다.

본 논문에서는 FCM 클러스터링을 이용하여 데이터를 분류하고 정량화 하여 지도학습을 진행하였다. 따라서 기존의 지도학습 방법에서 발생 할 수 있는 Over Fitting 문제를 최소화 하고 안정적인 학습을 진행 하는 것을 실험을 통해 확인하였다.

향후 연구과제로는 다양한 데이터 셋에 제안된 방법을 적용하고 데이터 셋에 효과적인 클러스터링 기법과 지도 학습 방법을 결합하는 학습 구조를 연구하여 방법의 성능을 개선할 것이다.

## References

[1] J. Singh and R. Banerjee, ‘‘A study on single and multi-layer perceptron neural network,’’ *Proceedings of 3rd Int. Conf. Comput. Methodologies Commun. (ICCMC)*, pp. 35-40, Mar. 2019.

[2] K. B. Kim, “Nucleus Recognition of Uterine Cervical Pap-Smears using FCM Clustering Algorithm,” *International Journal of Maritime Information and Communication Sciences*, vol. 6, no. 1, pp. 94-99, 2008.

[3] “Diabetes Dataset,” Kaggle, last modified 2020, accessed Jul. 2022, [Internet]. Available : https://www.kaggle.com/datasets/mathchi/diabetes-data-set