8.1 HCM 클러스터링 방법

HCM(Hard C-Means) 클러스터링 방법은 데이터 간의 거리를 기준으로 근접한 정도를 측정하고, 이를 바탕으로 데이터를 특성별로 분류하여 데이터들의 특성을 파악하는 데 사용됩니다. HCM은 0과 1, 즉 이진 논리에 의해 분리된 데이터가 그룹에 속해 있는지 아닌지를 판별합니다. HCM 클러스터링 분류 과정을 단계별로 나타내면 다음과 같습니다.

Step 1: 클러스터의 개수 $c(2 < c < n)$를 결정하고, 학습 데이터 수 m에 따라 소속행렬 U를 $U^{(r)} \in M_c$로 초기화합니다. 초기 소속행렬이므로 $r = 0$이고 $u_{ij}(i = 1, 2, \cdots, c$ 이고 $j = 1, 2, \cdots, m)$는 소속행렬 U의 파라미터입니다.

$$M_c = \left\{ U | u_{ij} \in \{0,\ 1\} \text{ 이고 } \sum_{i=1}^{c} u_{ij} = 1 \text{ 이고 } 0 < \sum_{j=1}^{m} u_{ij} < m \right\} \tag{8.1}$$

Step 2: 각 클러스터에 대한 중심벡터 v를 구합니다. i는 i 번째 클러스터, j는 j 번째 학습 데이터$(j = 1, \cdots, m)$, k는 학습 데이터의 k 번째 요소입니다.

$$v_{ik} = \frac{\sum_{j=1}^{m} u_{ij} \cdot x_{jk}}{\sum_{j=1}^{m} u_{ij}} \tag{8.2}$$

Step 3: 각 클러스터의 중심으로부터 데이터 간의 거리를 계산하여 새로운 소속행렬 $U^{(r+1)}$을 생성합니다. n은 학습 데이터의 크기(요소 개수)입니다.

$$d_{ij} = \|x_{jk} - v_{ik}\| = \left[\sum_{k=1}^{n} (x_{jk} - v_{ik})^2 \right]^{1/2} \tag{8.3}$$

$$u_{ij}^{(r+1)} = \begin{cases} 1 & d_{ij}^{(r)} = min \left\{ d_{ij}^{(r)}\ for\ all\ i \in c \right\} \\ 0 & otherwise \end{cases} \tag{8.4}$$

Step 4: 오류 허용 조건을 만족하면 학습을 종료하고, 그렇지 않다면 $r = r + 1$로 변경한 후 Step 2로 이동합니다.

$$\|U^{(r+1)} - U^{(r)}\| \leq \epsilon(\text{tolerance level}) \tag{8.5}$$

8.1.1 HCM 클러스터링의 수치적 예제

다음 주어진 데이터 샘플을 HCM 클러스터링 방법으로 분류해보겠습니다.

$$x_1 = \{1, 3\},\ x_2 = \{1.5, 3.2\},\ x_3 = \{1.3, 2.8\},\ x_4 = \{3, 1\}$$

Step 1: 초기 소속 행렬 $U^{(0)}$를 정의합니다.

$$U^{(0)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

Step 2: 각 클러스터에 대한 중심 벡터를 계산합니다.
주어진 예제에서 각 데이터는 두 개의 좌표로 이루어져 있으므로 $v_i = \{v_{i1}, v_{i2}\}$ 입니다. 따라서 첫 번째 클러스터($c = 1$)의 중심 벡터는 $v_1 = \{v_{11}, v_{12}\}$로 나타낼 수 있고, 두 번째 클러스터($c = 2$)의 중심벡터는 $v_2 = \{v_{21}, v_{22}\}$로 나타낼 수 있습니다.
식 8.2를 이용하여 중심 벡터 v를 계산하면,

$$v_{11} = \frac{(1)1 + (0)1.5 + (0)1.3 + (0)3}{1 + 0 + 0 + 0} = 1$$

$$v_{12} = \frac{(1)3 + (0)3.2 + (0)2.8 + (0)1}{1 + 0 + 0 + 0} = 3$$

$$v_{21} = \frac{(0)1 + (1)1.5 + (1)1.3 + (1)3}{0 + 1 + 1 + 1} = 1.93$$

$$v_{22} = \frac{(0)3 + (1)3.2 + (1)2.8 + (1)1}{0 + 1 + 1 + 1} = 2.33$$

따라서 $v_1 = \{v_{11}, v_{12}\} = \{1, 3\}$, $v_2 = \{v_{21}, v_{22}\} = \{1.93, 2.33\}$ 입니다.

Step 3: 각 클러스터 중심과 데이터 간의 거리를 계산하여 새로운 소속행렬 $U^{(r+1)}$을 생성합니다. 식 8.3을 이용하여 거리를 계산하면 다음과 같습니다.

첫 번째 클러스터 중심과의 거리 계산	두 번째 클러스터 중심과의 거리 계산
$d_{1j} = \left[(x_{j1} - v_{11})^2 + (x_{j2} - v_{12})^2\right]^{1/2}$	$d_{2j} = \left[(x_{j1} - v_{21})^2 + (x_{j2} - v_{22})^2\right]^{1/2}$
$d_{11} = \sqrt{(1-1)^2 + (3-3)^2} = 0$	$d_{21} = \sqrt{(1-1.93)^2 + (3-2.33)^2} = 1.15$
$d_{12} = \sqrt{(1.5-1)^2 + (3.2-3)^2} = 0.54$	$d_{22} = \sqrt{(1.5-1.93)^2 + (3.2-2.33)^2} = 0.97$
$d_{13} = \sqrt{(1.3-1)^2 + (2.8-3)^2} = 0.36$	$d_{23} = \sqrt{(1.3-1.93)^2 + (2.8-2.33)^2} = 0.79$
$d_{14} = \sqrt{(3-1)^2 + (1-3)^2} = 2.83$	$d_{24} = \sqrt{(3-1.93)^2 + (1-2.33)^2} = 1.71$

식 8.4를 이용하여 새로운 소속행렬을 계산하면 다음과 같습니다. 첫 번째 클러스터에 대한 계산과정만 나타내도록 하겠습니다.

$$min(d_{11}, d_{21}) = min(0, 1.15) = 0 \text{ 이므로 } u_{11} = 1$$
$$min(d_{12}, d_{22}) = min(0.54, 0.97) = 0.54 \text{ 이므로 } u_{12} = 1$$
$$min(d_{13}, d_{23}) = min(0.36, 0.79) = 0.36 \text{ 이므로 } u_{13} = 1$$
$$min(d_{14}, d_{24}) = min(2.83, 1.71) = 1.71 \text{ 이므로 } u_{14} = 0$$

각 클러스터와 데이터 간의 거리를 이용하여 새로운 소속행렬 $U^{(1)}$을 구하면

$$U^{(1)} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

가 됩니다.

Step 4: 초기 소속행렬 $U^{(0)}$와 새로운 소속행렬 $U^{(1)}$을 비교하여 학습 종료 조건(식 8.5)을 판단합니다. 학습 종료 조건을 만족하지 않으므로 $U^{(1)}$을 소속행렬로 설정하고 Step 2로 이동하여 다시 학습 과정을 진행합니다.

Step 5: 새로운 소속행렬 $U^{(1)}$을 이용하여 다시 중심 벡터 v를 계산하면,

$$v_{11} = \frac{(1)1 + (1)1.5 + (1)1.3 + (0)3}{1 + 1 + 1 + 0} = 1.26$$

$$v_{12} = \frac{(1)3 + (1)3.2 + (1)2.8 + (0)1}{1 + 1 + 1 + 0} = 3.0$$

$$v_{21} = \frac{(0)1 + (0)1.5 + (0)1.3 + (1)3}{0 + 0 + 0 + 1} = 3.0$$

$$v_{22} = \frac{(0)3 + (0)3.2 + (0)2.8 + (1)1}{0 + 0 + 0 + 1} = 1.0$$

따라서 $v_1 = \{v_{11}, v_{12}\} = \{1.26, 3.0\}$, $v_2 = \{v_{21}, v_{22}\} = \{3.0, 1.0\}$ 입니다.

새롭게 계산한 각 클러스터 중심과 데이터 간의 거리를 계산하면 다음과 같습니다.

첫 번째 클러스터 중심과의 거리 계산	두 번째 클러스터 중심과의 거리 계산
$d_{11} = \sqrt{(1-1.26)^2 + (3-3)^2} = 0.26$	$d_{21} = \sqrt{(1-3)^2 + (3-1)^2} = 2.83$
$d_{12} = \sqrt{(1.5-1.26)^2 + (3.2-3)^2} = 0.31$	$d_{22} = \sqrt{(1.5-3)^2 + (3.2-1)^2} = 2.66$
$d_{13} = \sqrt{(1.3-1.26)^2 + (2.8-3)^2} = 0.20$	$d_{23} = \sqrt{(1.3-3)^2 + (2.8-1)^2} = 2.48$
$d_{14} = \sqrt{(3-1.26)^2 + (1-3)^2} = 2.65$	$d_{24} = \sqrt{(3-3)^2 + (1-1)^2} = 0.0$

새로운 소속행렬을 계산하면 다음과 같습니다. 첫번째 클러스터에 대한 계산과정만 나타내도록 하겠습니다.

$$min(d_{11}, d_{21}) = min(0.26, 2.83) = 0.26 \text{ 이므로 } u_{11} = 1$$
$$min(d_{12}, d_{22}) = min(0.31, 2.66) = 0.31 \text{ 이므로 } u_{12} = 1$$
$$min(d_{13}, d_{23}) = min(0.20, 2.48) = 0.20 \text{ 이므로 } u_{13} = 1$$
$$min(d_{14}, d_{24}) = min(2.65, 0) = 0 \text{ 이므로 } u_{14} = 0$$

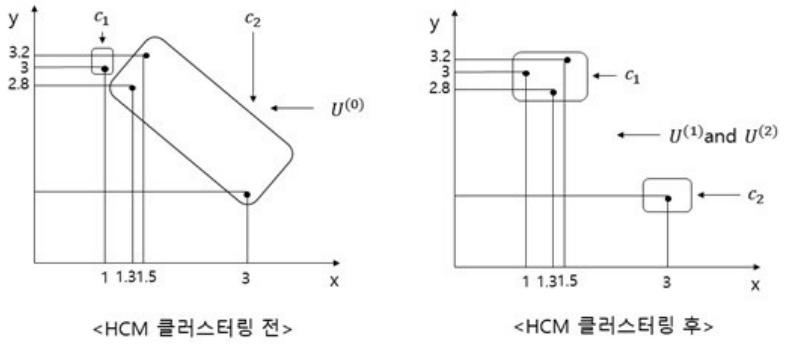
각 클러스터와 데이터 간의 거리를 이용하여 새로운 소속행렬 $U^{(2)}$를 구하면

$$U^{(2)} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

가 됩니다.

소속행렬 $U^{(1)}$와 소속행렬 $U^{(2)}$를 비교하여 학습 종료 조건(식 8.5)을 판단합니다. 학습 종료 조건을 만족하므로 학습을 종료합니다. 최종 소속행렬과 중심값은 다음과 같습니다.

$$U = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$
$$v_1 = \{v_{11}, v_{12}\} = \{1.26, 3.0\}\,,\ v_2 = \{v_{21}, v_{22}\} = \{3.0, 1.0\}$$



〈그림 8.2〉 HCM 클러스터링 방법을 이용한 데이터 분류

8.1.2 C# HCM 소스 프로그램

코드 8.1: HCM 클러스터링 방법을 이용한 분류

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.IO;

class HCM
```