K-Means 군집분석 실습 보고서

Y2023011 윤요섭

Ⅰ. 자료 및 분석 방법

본 보고서는 bankloan.sav 데이터를 이용하여 IBM SPSS Statistics를 통해 Multiclass ANN/SVM 실습과제를 수행하였다. bankloan.sav 데이터 변수와 이에 대한 설명은 다음과 같다.

age	Age in years
employ	Years with current employer
address	Years at current address
debtinc	Debt to income ratio (×100)
creddebt	Credit card debt in thousands
othdebt	Other debt in thousands
default	Previously defaulted

군집 분석은 거리가 가까운 (유사도가 높은) 개체들을 서로 묶어 하나의 그룹으로 정리하는 분석 기법이다. 군집화 방법의 2가지 유형으로는 군집의 개수를 사후적으로 결정 할 수 있는 계층적 군집화 방법과 미리 군집의 개수를 정하고 시작해야 하는 비계층적 군집화 방법이 있다.

본 보고서에는 비계층적 군집화 방법 중 하나인 K-Means를 활용아혀 bankloan.sav에 저장된 표본들을 4개의 군집으로 군집화 하였다.

K-Means는 사전에 정해진 군집의 개수(K)에 따라 데이터를 군집으로 그룹화 하는 비지도 학습 알고리즘으로, 가장 단순하고 손쉽게 적용할 수 있는 군집분석 알고리즘이다.

Ⅱ. 분석 과정과 결과 및 해석

본 실습에서는 데이터들을 4개의 집단으로 군집화하기 위해, 군집 수를 4로 입력하였다. 또한, 초기에 임의로 4개의 군집 중심점(임시)를 설정한 뒤, 군집화를 수행하고 각 군집별 중심점을 계산하여, 새로운 중심점을 기준으로 군집화를 수행하는 반복계산 수를 최대 20으로 하였다.

군집 분석 결과는 다음 장과 같다.

Ⅱ-가 초기 군집중심 결과

초기 군집중심

	군집			
	1	2	3	4
Age in years	51	54	20	33
Years with current employer	10	18	0	14
Years at current address	1	34	1	8
Debt to income ratio (x100)	13.20	8.50	2.30	41.30
Credit card debt in thousands	4.96	3.29	.04	15.02
Other debt in thousands	.85	6.40	.35	14.72
Previously defaulted	1	0	0	1

초기 군집 중심 결과를 토대로 본 4개의 군집의 특성은 다음과 같다.

1. 1군집, 2군집의 경우, 높은 연령대와 많은 근속기간의 특성을 가지고 있다.
2. 1군집과 4군집의 경우, 부채 대 소득 비율이 다른 군집보다 높은 특성을 가지고 있다.
3. 2군집과 4군집의 경우, 신용카드 부채가 다른 군집보다 높은 특성을 가지고 있다.
4. 2군집과 4군집의 경우, 기타 부채가 다른 군집보다 높은 특성을 가지고 있다.

군집화를 수행하고 각 군집별 중심점을 계산하여, 새로운 중심점을 기준으로 군집화를 수행하는 반복계산하다보면, 일부 좌표의 소속 군집이 변화한다. 실습에서 군집중심의 변화량 정보가 들어있는 반복계산정보는 아래와 같다.

반복계산정보[a]

반복계산	군집중심의 변화량			
	1	2	3	4
1	14.545	14.398	12.193	16.879
2	1.488	1.592	.472	4.289
3	1.030	.962	.383	2.129
4	.454	.883	.251	1.290
5	.383	1.034	.251	1.291
6	.296	.473	.222	.587
7	.166	.152	.129	.260
8	.126	.000	.093	.311
9	.000	.000	.000	.000

a. 군집 중심값의 변화가 없거나 작아 수렴이
일어났습니다. 모든 중심에 대한 최대 절대 좌표
변경은 .000입니다. 현재 반복계산은 9입니다. 초기 중심
간의 최소 거리는 34.716입니다.

반복계산 정보를 보면, 8번 반복할 때까지 각 군집의 변화가 이루어진 것을 볼 수 있다. 이는 군집의 중심점이 변하면서 소속 군집이 변한 것인데, 4군집의 경우 가장 많은 변화량을 보인 것을 알 수 있다.

반복계산 이후, 최종 군집중심 결과는 아래와 같다.

ㅍ-나 최종 군집중심 결과

최종 군집중심

	군집			
	1	2	3	4
Age in years	39	46	27	34
Years with current employer	10	17	4	7
Years at current address	9	20	4	8
Debt to income ratio (x100)	7.09	10.92	8.03	21.10
Credit card debt in thousands	1.10	3.43	.66	2.90
Other debt in thousands	2.28	5.72	1.42	6.06
Previously defaulted	0	0	0	1

최종 군집중심을 토대로 본 4개의 군집의 특성은 다음과 같다.

1. 1군집, 2군집의 경우, 높은 연령대와 많은 근속기간의 특성을 가지고 있다.
2. 2군집과 4군집의 경우, 부채 대 소득 비율이 다른 군집보다 높은 특성을 가지고 있다.
3. 2군집과 4군집의 경우, 신용카드 부채가 다른 군집보다 높은 특성을 가지고 있다.
4. 2군집과 4군집의 경우, 기타 부채가 다른 군집보다 높은 특성을 가지고 있다.

초기 군집중심과 비교하여 볼 때, 1번의 경우는 마찬가지로 1군집과 2군집이 높은 연령대와 많은 근속기간의 특성을 가지고 있다는 것을 알 수 있으나, 그 중심점이 더 낮아진 것을 알 수 있다.

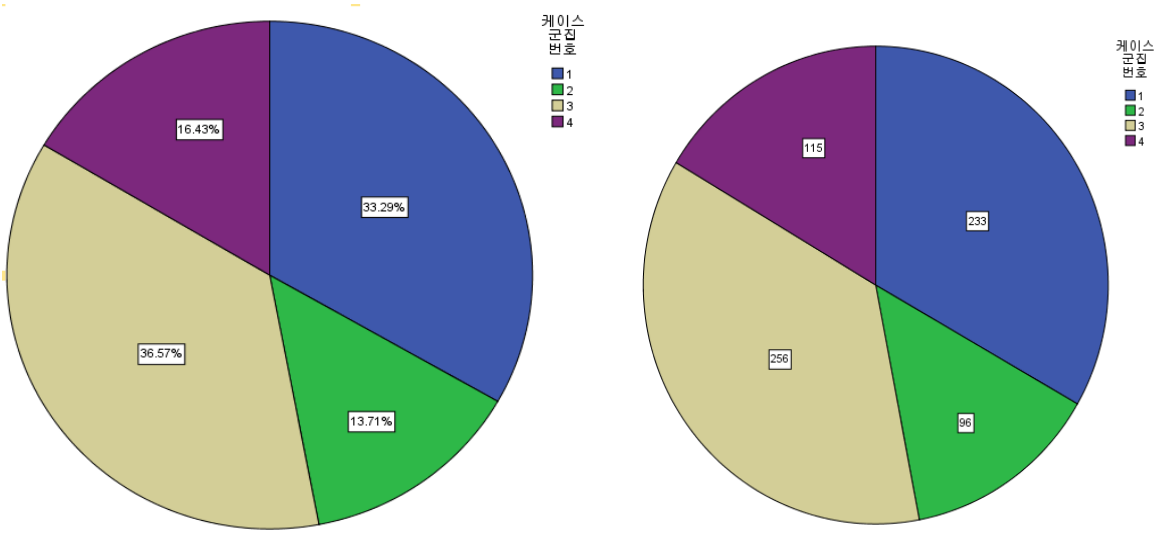
부체 대 소득 비율의 경우, 초기 군집중심에서는 1,4군집이 다른 군집보다 높은 특성을 가지고 있었으나, 최종 군집 중심에서는 2,4군집이 다른 군집보다 높은 특성을 가지고 있는 것으로 변화하였다.

마지막으로, 기타 부채의 경우도 2,4 군집이 다른 군집보다 높은 특성을 가지고 있는 것을 확인할 수 있었으나, 초기 군집 중심에 비해서 군집 중심점이 보다 더 작아진 것을 확인 할 수 있다.

최종 군집의 각 비율과 레코드 수는 아래와 같다.

ㅍ-다 최종 군집중심 군집별 비율과 빈도

[군집별 비율과 빈도]



3군집의 경우, 256(36.57%)로 가장 많은 군집에 해당한다. 1군집의 경우, 233(33.29%)로 두 번째로, 많은 군집을 가지고 있는 것을 확인 할 수 있다.

4군집은 115(16.43%)로 세 번째로 많은 군집을 가지고 있으며, 마지막으로 2군집의 경우, 96(13.71%)로 가장 적은 군집에 해당하는 것을 알 수 있다.