혼합형 데이터에 대한 군집분석 알고리즘 비교 및 사례분석*

Jeong-Min Lee[1], So-Yeon Jo[1], Kyung-Joon Cha[2]

[1] Department of Applied Statistics, Hanyang University, Korea
[2] Department of Mathematics and
Research Institute for Convergence of Basic Science, Hanyang University, Korea

Introduction

군집분석은 유사성이 높은 데이터를 같은 군집으로 그룹화하는데 사용되는 비지도 머신러닝 기법이다. 많은 분야에서 연속형 데이터와 범주형 데이터가 혼합되어 있기 때문에, 군집분석에 주로 사용되는 K-means 군집화 방법은 범주형 데이터의 정보를 손실할 수 있다. 따라서 혼합형 데이터를 다룰 수 있는 적절한 군집화 방법을 찾는 것이 중요하다. 본 연구에서는 H사이버대학교의 2020년 학업실태조사 데이터를 Gower 거리를 이용한 군집화 방법, K-prototype 군집화 방법, KAMILA 군집화 방법을 이용하여 비교 분석하였다.

Mixed data clustering algorithms

- Gower Distance[1] : Gower의 방법은 범주형 변수에 대해서는 단순 대응 방법을, 연속형 변수는 최소-최대 정규화 방법을 이용하여 거리를 계산한다. 이를 통해 모든 변수가 가지는 거리의 범위를 동일하게 하여, 개체 간 거리계산에 변수들이 균등한 영향을 주도록 한다. 최종적으로 맨해튼(Manhattan) 거리 측도를 이용하여 계산한다.

$$d(X,Y) = \sum_{j=1}^{q} |x_j - y_j| + \sum_{j=q+1}^{p} s(x_j y_j)$$

- K-prototype clustering[2] : K-prototype 군집화 방법은 K개의 그룹의 중심을 정의하며, 연속형 변수의 경우에는 평균으로, 범주형 변수의 경우에는 Modes로 정의한다. 연속형 변수에 대해서는 유클리디안(Euclidean) 거리를 이용하고, 범주형 변수에 대해서는 해밍(Hamming) 거리를 이용한다.

$$d_2(X,Y) = \sum_{j=1}^{q} (x_j - y_j)^2 + \gamma \sum_{j=q+1}^{p} \delta(x_j y_j)$$

- KAMILA clustering[3] : KAMILA 군집화 방법은 K-means 군집화 방법을 기반으로 하였다. 연속형 변수에 대한 중심 집합과 범주형 변수에 대한 모수 집합으로 시작하여, 연속형 변수의 경우에는 가장 가까운 중심을 가진 유클리디안 거리를 계산하고 커널 함수를 통해 연속형 변수의 혼합 분포를 추정한다. 범주형 변수의 경우 주어진 데이터를 관측할 확률을 계산하여, 두 성분의 합계에 대한 로그 우도(likelihood)를 사용해 각 개체에 적합한 군집을 찾는다.

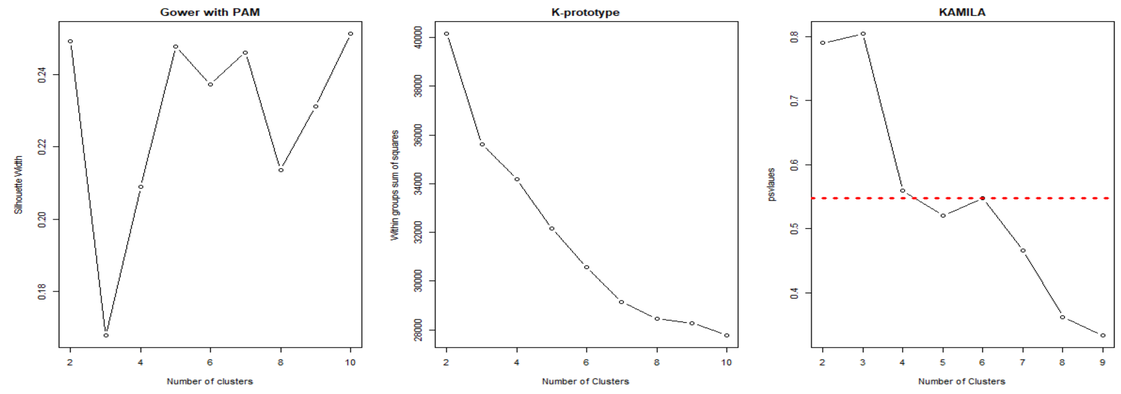
$$H_i^{(t)}(g) = log\left[\hat{f}_V^{(t)}\left(d_{ig}^{(t)}\right)\right] + log\left[c_{ig}^{(t)}\right]$$

Data description

- Data description : 2020년 2학기에 H사이버대학의 1~4학년 4095명을 대상으로 학업실태조사를 진행하였다. 분석에 쓰인 데이터는 5점 리커트 척도의 설문조사 데이터와 인적사항 데이터로 구성되었다.
- 인적사항 : 학부, 학년, 성별의 3개의 범주형 데이터와 나이, 신청학점, 학점취득비율, 정규평점의 4개의 연속형 데이터로 구성된다.
- 설문조사 : 56개의 설문조사 문항을 요인분석을 통해 학업참여동기, 학업역량, 학업여건, 교육 만족도, 학업수행, 성취감의 6개의 대분류로 분류하였다. 세부문항들의 평균으로 대분류한 설문조사 데이터를 구간화 하여 5점 척도를 가지는 범주형 데이터로 변환하였다.

Results

1. 최적의 군집 개수(K) 찾기



- 각 군집화 방법 별 최적의 군집 개수를 찾는 방법이 다르다.
- Gower 거리를 이용한 방법은 Silhouette 거리가 비슷하게 나타나는 군집이 있고, K-prototype 군집화 방법은 Elbow 방법을 써서 확인한 결과, 기울기가 바뀌는 점이 여러 개 나타나 최적의 군집 개수를 찾을 수 없다.
- KAMILA 군집화 방법의 경우에는 Ps값이 0.5 이상에서 최적의 군집 수가 나타나기 때문에, 이 데이터의 경우에는 군집의 수가 6개 일 때 군집분석이 최적화 되었다고 할 수 있다.

2. KAMILA 군집화 결과

연속형 변수	1 median (N=540)	2 median (N=790)	3 median (N=372)	4 median (N=1309)	5 median (N=978)	6 median (N=106)	Kruskal-wallis test p-value
나이	**39.5 (9.17)**	28 (7.34)	29 (8.66)	*27 (5.20)*	**49 (6.59)**	28.5 (9.18)	< 2.2e-16
신청학점	*12 (3.06)*	**18 (3.41)**	15 (4.21)	**18 (2.95)**	**18 (2.74)**	15 (5.57)	< 2.2e-16
학점취득비율	1.00 (0.04)	1.00 (0.04)	*0.89 (0.13)*	1.00 (0.02)	1.00 (0.02)	*0.17 (0.20)*	< 2.2e-16
정규평점	3.75 (0.56)	3.75 (0.56)	*2.25 (0.67)*	3.75 (0.48)	3.70 (0.52)	*0.00 (0.52)*	< 2.2e-16

범주형 변수		1 frequency (N=540)	2 frequency (N=790)	3 frequency (N=372)	4 frequency (N=1309)	5 frequency (N=978)	6 frequency (N=106)	Row Sum (Column %)	Pearson's Chi-squared test p-value
학부	공학	*120 (22.2%)*	**296 (37.5%)**	98 (26.3%)	**469 (35.8%)**	*195 (19.9%)*	**35 (33.0%)**	1213 (29.6%)	< 2.2e-16
	디자인	**111 (20.6%)**	98 (12.4%)	55 (14.8%)	154 (11.8%)	134 (13.7%)	**21 (19.8%)**	573 (14.0%)	
	인문	309 (57.2%)	*396 (50.1%)*	219 (58.9%)	686 (52.4%)	**649 (66.4%)**	*50 (47.2%)*	2309 (56.4%)	
학년	1	*40 (7.4%)*	154 (19.5%)	**105 (28.2%)**	335 (25.6%)	233 (23.8%)	**28 (26.4%)**	895 (21.9%)	< 2.2e-16
	2	*24 (4.4%)*	106 (13.4%)	63 (16.9%)	193 (14.7%)	**191 (19.5%)**	**23 (21.7%)**	600 (14.7%)	
	3	**119 (22.0%)**	316 (40.0%)	*93 (25.0%)*	**575 (43.9%)**	373 (38.1%)	*27 (25.5%)*	1503 (36.7%)	
	4	**357 (66.1%)**	214 (27.1%)	111 (29.8%)	*206 (15.7%)*	*181 (18.5%)*	28 (26.4%)	1097 (26.8%)	
성별	남	213 (39.4%)	**409 (51.8%)**	**188 (50.5%)**	511 (39.0%)	379 (38.8%)	*40 (37.7%)*	1740 (42.5%)	3.84e-10
	여	327 (60.6%)	*381 (48.2%)*	184 (49.5%)	798 (61.0%)	599 (61.2%)	**66 (62.3%)**	2355 (57.5%)	

설문조사	1 mean (N=540)	2 mean (N=790)	3 mean (N=372)	4 mean (N=1309)	5 mean (N=978)	6 mean (N=106)
학업참여동기	3.86 (0.48)	**4.27 (0.71)**	3.69 (0.72)	*3.52 (0.57)*	3.87 (0.57)	3.67 (0.78)
학업역량	3.58 (0.56)	**4.18 (0.67)**	3.34 (0.70)	3.37 (0.57)	3.39 (0.56)	3.37 (0.79)
학업여건	3.00 (0.36)	3.24 (0.60)	*2.88 (0.46)*	*2.89 (0.39)*	3.00 (0.34)	*2.86 (0.52)*
교육만족도	3.21 (0.57)	**3.75 (0.70)**	3.00 (0.60)	*2.96 (0.40)*	3.23 (0.49)	3.11 (0.73)
학업수행	2.86 (0.66)	**3.49 (0.86)**	*2.40 (0.82)*	2.66 (0.59)	2.94 (0.60)	*2.22 (0.81)*
성취감	2.90 (0.66)	**3.43 (0.79)**	*2.49 (0.77)*	2.65 (0.57)	2.93 (0.66)	*2.49 (0.80)*

Conclusion

- 6개의 군집 중에서 2번 집단의 경우 18학점 정도의 학점을 신청하고, 성적도 우수하다. 또한, 학업 여건을 제외하고 나머지 대분류에서 높은 만족도를 확인할 수 있다.
- 반대로, 6번 집단의 경우 학점취득비율과 정규평점이 낮게 나타나며, 다른 군집에 비해 1,2학년에 집중되어 있는 것을 알 수 있다. 또한, 학업여건에서 가장 낮은 만족도를 보이는 것을 알 수 있다.
- 추후에 다른 데이터를 통해서 다양한 방법을 비교 분석할 것이다.

References

[1] Gower, J. C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 857-871.

[2] Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 2(3), 283-304.

[3] Foss, A., Markatou, M., Ray, B., & Heching, A. (2016). A semiparametric method for clustering mixed data. Machine Learning, 105(3), 419-458.

* 본 연구는 기초과학융합연구소(Research Institute for Convergence of Basic Science, NRF-2020R1A6A1A06046728)의 지원으로 이루어졌습니다.