K-means Clustering for Environmental Indicator Survey Data

Hee-Chang Park[1], Kwang-Hyun Cho[2]

Abstract

There are many data mining techniques such as association rule, decision tree, neural network analysis, clustering, genetic algorithm, bayesian network, memory-based reasoning, etc. We analyze 2003 Gyeongnam social indicator survey data using k-means clustering technique for environmental information. Clustering is the process of grouping the data into clusters so that objects within a cluster have high similarity in comparison to one another. In this paper, we used k-means clustering of several clustering techniques. The k-means clustering is classified as a partitional clustering method. We can apply k-means clustering outputs to environmental preservation and environmental improvement.

keywords : clustering, data mining, k-means clustering, environmental indicator survey

1. 서론

데이터마이닝이란 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정으로 숨겨진 지식, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 바탕으로 의사결정 등을 위한 정보로 활용하고자 하는 것이다.

데이터마이닝 기법으로는 군집분석(cluster analysis), 연결 분석(link analysis), 판별분석(discrimination analysis), 연관성규칙(association rule), 의사결정나무(decision tree) 기법, 신경망모형(neural network) 등의 분석 기법이 있다. 본 논문에서 적용한 k-평균 클러스터링은 데이터들을 k개의 군집으로 임의로 분할을 하여 군집의 평균을 대표값으로 분할해 나가는 방법으로 데이터들을 유사성을 바탕으로 재배치를 하는 방법이다.

k-평균 클러스터링은 MacQueen(1967)에 의해 처음 소개되었다. Kaufman과 Rousseeuw(1990)는 k-means 알고리즘이 이상값에 민감한 것을 보완하여 군집의 대표값을 중앙값으로 하는 k-medoids 방법인 PAM(Partitioning Around Medoids)을 제안하였다. PAM은 적은 데이터 셋에서는 좋은 결과를 보였으나 많은 양의 데이터 셋

1) First author : Professor, Department of Statistics, Changwon National University, Changwon, Gyungnam, 641-773, Korea
E-mail : hcpark@sarim.changwon.ac.kr
2) Graduate Student, Department of Statistics, Changwon National University, Changwon, Gyungnam, 641-773, Korea

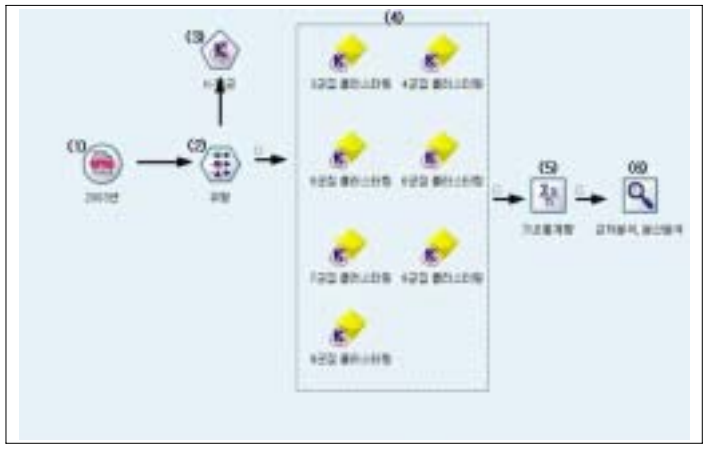
에서는 효과적이지 못하다. 그래서 이들은 많은 양의 데이터를 취급하기 위해 CLARA(Clustering LARge Applications) 알고리즘을 제안하였다. CLARA는 데이터를 샘플링하여 PAM을 적용한 방법이다. 표본을 잘 뽑았다면 표본의 중앙값은 전체 데이터의 중앙값에 근사한다. 더 나은 근사값을 위해 CLARA는 다중 샘플을 사용한다. Ng와 Han(1994) 그리고 Ester 등(1999)은 CLARA를 더욱 향상시킨 CLARANS(Clustering Large Applications based on RANdomized Search)를 제안하였으며 Huang(1997)은 k-means가 연속형 데이터에 대해 한정된 단점을 보완한 연속형과 범주형의 혼합된 데이터에 대한 k-prototypes 알고리즘을 제시하였다.

Chu, Roddick 와 Pan 등(2002)은 MCMRS(Multi-Centroid, Multi-Run Sampling Scheme) 알고리즘을 제시하였으며, Zaki 등(1988)은 각 군집에서 입력 벡터의 ensemble average가 계산되어지는 새로운 비모수 분류 처리인 EA(Ensemble Average) 알고리즘을 제시하였다.

본 논문에서는 2003년 경상남도에서 조사한 사회지표 조사 자료에 대하여 k-평균 클러스터링 기법 적용을 연구하고자 한다. k-평균 클러스터링 적용 시, 군집의 수를 3~9개로 설정하여 각 모형을 생성하고 생성된 모형을 탐색 한 후 군집의 속성이 가장 잘 파악되는 군집의 개수를 선택하여 도민의 환경의식의 심층적 파악을 위하여 각 군집에 대한 환경관련 문항의 속성 분석을 실시한다. 본 논문의 2절에서는 k-평균 클러스터링 적용에 대하여 기술하고 3절에서는 3~9개의 k-평균 클러스터링의 결과를 비교하다. 4절에서는 k-평균 클러스터링을 이용한 자료 분석 결과를 기술한 후, 5절에서 결론을 맺는다.

2. k- 평균 클러스터링 적용

Clementine 10.0을 이용한 사회지표 조사 자료에 대한 k-평균 클러스터링 스트림은 <그림 1>과 같다.



<그림 1> k-평균 클러스터링 적용

[단계1] 자료 선택
k-평균 클러스터링에 적용할 자료를 선택한다. 본 논문에서 사용한 자료는 2003년 경상남도에 조사한 사회 지표조사에 대한 데이터이다. 자료의 구조는 크게 일반사항(인구통계학적 문항)과 도민의식조사부문으로 나누어져 있다. 일반사항은 조사응답자의 연령, 성별, 학력, 가구주와의 관계, 결혼유무, 직업으로 구성되어 있으며, 도민의식조사부문은 소득 소비, 보건체육, 고용노사, 교육, 환경교통, 사회, 정보화, 문화여가, 안전의 9개 분야로 되어 있고 총 41문항으로 구성되어 있다.

[단계2] 유형(자료의 선정)
k-평균 클러스터링에 사용할 문항을 선택한다. k-평균 클러스터링 분석을 위하여 연령, 주관적 사회계층, 학력 문항을 선정하고 각 군집의 속성 분석을 위하여 자녀의 환경오염 저감 행동 유무, 1회용품 사용빈도, 환경정책 시급과제, 산림의 중요성 체감, 우선적 의식 개선 부문, 최우선 행정 분야의 환경관련 문항을 선정한다.

[단계3] k-평균 클러스터링
k-평균 클러스터링을 실행한다. 이 때 클러스터링에 사용할 입력변수는 연령, 주관적 사회계층, 학력을 선택하고 군집의 수를 3~9로 지정하여 총 7개의 k-평균 클러스터링 결과를 산출한다.

[단계 4] k-평균 클러스터링 결과 분석 및 비교
3~9개의 k-평균 클러스터링에 대하여 군집의 결과를 분석하고 각 k-평균 클러스터링의 결과를 비교하여 군집의 특성이 가장 명확하게 구분되는 군집의 수를 결정한다. 7개의 k-평균 클러스터링의 결과 분석 및 비교는 3장에서 다루도록 한다.

[단계 5] 환경관련문항 자료 탐색(기술통계량)
환경관련 문항에 대하여 자료를 탐색한다. 각 군집별 환경관련 문항의 속성을 파악하기 전에 원 자료에 대한 자료를 탐색한다.

[단계 6] 각 군집별 교차분석 및 판별분석
각 군집에 대한 환경 관련 문항에 대하여 차이가 있는지 분석하기 위하여 이산형 문항에 대해서는 교차표에 의한 카이제곱 검정을 실시하고 연속형 문항에 대해서는 분산분석을 실시한다.

3. k-평균 클러스터링의 결과 비교

3~9개의 군집으로 k-평균 클러스터링의 결과를 비교한다. 각 군집의 k-평균 클러스터링의 결과는 <표 1> ~ <표 7>과 같다.

<표 1> 3군집 k-평균 클러스터링 결과

항목 \ 군집	군집-1	군집-2	군집-3
주관적 사회계층	1.646	3.117	2.587
연령	49.287	45.693	32.35
학력	3.049	3.403	7.09
군집 레코드 수	4738	2655	2607

<표 2> 4군집 k-평균 클러스터링 결과

항목 \ 군집	군집-1	군집-2	군집-3	군집-4
주관적 사회계층	1.495	3.174	1.797	2.646
연령	57.858	33.061	31.298	41.774
학력	2.405	7.219	6.747	3.666
군집 레코드 수	2928	1463	1233	4376

<표 3> 5군집 k-평균 클러스터링 결과

항목 \ 군집	군집-1	군집-2	군집-3	군집-4	군집-5
주관적 사회계층	1.777	3.133	1.837	2.305	4.176
연령	62.943	33.773	32.353	38.849	41.963
학력	2.033	7.284	7.058	3.857	3.524
군집 레코드 수	2660	1421	1082	4570	267

<표 4> 6군집 k-평균 클러스터링 결과

항목 \ 군집	군집-1	군집-2	군집-3	군집-4	군집-5	군집-6
주관적 사회계층	1.655	4.195	1.851	1.717	3.092	3.0
연령	65.132	32.13	32.441	38.797	44.584	32.751
학력	1.834	6.965	7.091	3.833	3.497	7.138
군집 레코드 수	2017	231	1065	2922	2464	1301

<표 5> 7군집 k-평균 클러스터링 결과

항목 \ 군집	군집-1	군집-2	군집-3	군집-4	군집-5	군집-6	군집-7
주관적 사회계층	1.544	4.196	1.851	3.067	3.115	3.0	1.724
연령	64.457	32.074	32.441	36.11	60.692	33.828	38.656
학력	1.897	6.974	7.091	3.949	2.524	7.274	3.839
군집 레코드 수	1902	230	1065	1693	993	1223	2894

<표 6> 8군집 k-평균 클러스터링 결과

항목 \ 군집	군집-1	군집-2	군집-3	군집-4	군집-5	군집-6	군집-7	군집-8
주관적 사회계층	1.663	4.148	1.0	3.0	4.114	2.612	1.712	2.545
연령	65.265	31.681	31.052	44.256	47.843	37.815	39.417	27.443
학력	1.821	7.162	6.844	3.522	3.266	8.064	3.799	6.184
군집 레코드 수	2001	216	173	2232	229	1055	2871	1223

<표 7> 9군집 k-평균 클러스터링 결과

항목 \ 군집	군집-1	군집-2	군집-3	군집-4	군집-5	군집-6	군집-7	군집-8	군집-9
주관적 사회계층	1.543	4.077	1.0	3.0	4.212	2.613	1.722	2.539	3.07
연령	64.494	36.884	31.052	37.59	36.059	37.815	39.225	27.388	61.6
학력	1.889	7.58	6.844	3.895	3.901	8.065	3.81	6.2	2.425
군집 레코드 수	1892	181	173	1567	203	1054	2831	1208	891

3~9개의 군집으로 k-평균 클러스터링의 결과, 4~9개의 k-평균 클러스터링 결과는 각 군집을 명확하게 구분하기가 쉽지 않다. 반면, 3군집 k-평균 클러스터링의 결과는 군집의 특성이 명확하게 구분된다. 군집-1의 집단은 다른 집단들에 비하여 연령이 높고 학력이 낮으며 주관적 사회계층이 낮은 집단으로 분류되고 군집-2는 다른 집단들에 비하여 상대적으로 연령과 학력은 보통이고 주관적 사회 계층은 높은 성향을 가지으며 군집-3의 집단은 다른 집단들에 비하여 상대적으로 연령은 낮고 학력은 높으며 주관적 사회계층이 보통의 성향을 가지는 집단으로 분류되었다. 이와 같이 3개의 군집으로 k-평균 클러스터링 한 결과가 각 군집의 특성을 가장 잘 분류되어 3군집 k-평균 클러스터링의 결과를 바탕으로 속성을 분석한다.

4. 자료 분석 결과

2003년 조사된 경남 사회지표조사 자료에서 도민들의 환경의식을 종합적으로 파악하여 그 현안을 보다 심층적으로 분석하기 위하여 각 군집과 환경관련 문항의 응답과의 차이를 파악한다. 분석을 위하여 카이제곱 검정을 실시하였다. 카이제곱 검정의 결과 각 군집별로 모든 환경관련 문항과의 차이가 있다고 나타났다(유의수준 0.05). <표 8> ~ <표 13>은 각 군집별 환경관련 문항과의 교차표이다.

<표 8> 각 군집과 자녀의 환경 저감 행동 유무와의 교차표

군집 \ 환경오염 저감행동			자녀의 환경오염 저감행동 경험 유무		
			없음	모름	있음
군집	1	빈도	526	1473	2727
		수정된 잔차	-1.3	**8.7**	-7.1
	2	빈도	322	659	1653
		수정된 잔차	1.2	-2.8	1.7
	3	빈도	302	560	1716
		수정된 잔차	.3	-7.1	**6.3**
전체	빈도		1150	2692	6096

군집 1의 집단은 자녀의 환경 저감 행동 유무에 대하여 모름의 응답 비율이 높고 군집 3의 집단은 있음의 응답 비율이 높다.

<표 9> 각 군집과 1회용품 사용빈도와의 교차표

군집 \ 1회용품 사용정도			1회용품의 사용정도			
			거의 사용 않음	가끔 사용	많이 사용	아주 많이 사용
군집	1	빈도	2225	2319	172	22
		수정된 잔차	**14.4**	-9.8	-7.1	-5.7
	2	빈도	971	1517	128	38
		수정된 잔차	-3.6	**3.7**	-1.3	2.0
	3	빈도	756	1570	232	49
		수정된 잔차	-12.8	**7.3**	**9.5**	**4.5**
전체	빈도		3952	5406	532	109

군집 1의 집단은 1회 용품 사용 빈도에 대하여 거의 사용하지 않음의 응답 비율이 높고 군집 2의 집단은 가끔 사용의 응답 비율이 높으며 군집 3의 집단은 가끔 사용, 많이 사용, 아주 많이 사용의 응답 비율이 높다.

<표 10> 각 군집과 환경정책 시급과제와의 교차표

군집 \ 환경오염 과제			환경오염 완화방법					
			단속 공무원 증원	주민의 환경의식 개혁	환경사범 벌칙 강화	환경보호시설 확충	민간환경단체 강화	기타
군집	1	빈도	453	2465	607	940	155	114
		수정된 잔차	**3.5**	-3.9	**2.3**	-1.5	.4	**5.6**
	2	빈도	234	1448	295	570	79	27
		수정된 잔차	.6	.6	-1.7	1.5	-.7	-3.0
	3	빈도	166	1495	300	536	85	24
		수정된 잔차	-4.6	**3.9**	-.9	.1	.2	-3.4
전체	빈도		853	5408	1202	2046	319	165

군집 1의 집단은 환경 정책 시급과제로서 단속 공무원 증원, 환경사범벌칙 강화, 기타의 응답 비율이 높고 군집 3의 집단은 주민의 환경의식 개혁이 가장 응답 비율이 높다.

<표 11> 각 군집과 산림의 중요성 체감과의 교차표

군집 \ 산림 중요성			산림 중요성의 체감도				
			느끼지 못함	그저 그럼	모르겠음	느낌	매우 느낌
군집	1	빈도	96	913	303	2166	1257
		수정된 잔차	.0	**4.9**	**9.3**	-1.3	-6.8
	2	빈도	63	457	99	1274	761
		수정된 잔차	1.5	-.2	-1.9	1.9	-1.5
	3	빈도	43	362	36	1203	962
		수정된 잔차	-1.6	-5.4	-8.7	-.3	**9.2**
전체	빈도		202	1732	438	4643	2980

군집 1의 집단은 산림 중요성의 체감에 대하여 그저 그럼, 모르겠음의 응답비율이 높고 군집 3의 집단은 매우 느낌의 응답 비율이 높다.

<표 12> 각 군집과 우선적 의식 개선 부문과의 교차표

군집 \ 우선과제			선진경남 건설을 위한 도민의식 최우선 개선과제							
			기초질서의식	교육의식	소비의식	환경의식	정치의식	근로의식	이웃과의 유대의식	기타
군집	1	빈도	2648	301	415	498	230	75	403	163
		수정된 잔차	-2.9	-4.9	**4.5**	-.8	-.2	**1.8**	**3.6**	**4.1**
	2	빈도	1453	242	219	305	150	33	189	59
		수정된 잔차	-3.2	**3.2**	1.7	1.4	**2.1**	-.6	-.9	-1.9
	3	빈도	1638	229	118	273	111	28	158	51
		수정된 잔차	**6.5**	**2.4**	-6.8	-.6	-1.8	-1.5	-3.3	-2.8
전체	빈도		5739	772	752	1076	491	136	750	273

군집 1의 집단은 우선적 이식 개선 부문에 대하여 소비의식, 이웃과의 유대의식, 기타의 응답 비율이 높고 군집 2의 집단은 교육의식, 정치의식의 응답비율이 높으며 군집 3의 집단은 기초질서의식, 교육의식의 응답 비율이 높다.

<표 13> 각 군집과 최우선 행정 분야와의 교차표

			지방자치단체장의 최우선 행정 추진분야									
			환경 보존	도민 복지	주민소득 증대	교통 개선	지역문화역사 정립	치안유지	교육개선	행정서비스 개선	주민의식개선 및 화합	기타
군집	1	빈도	841	1376	1402	420	97	109	228	149	109	5
		수정된 잔차	-4.7	.4	**10.6**	-2.4	-1.7	-2.2	-3.7	-2.5	-1.3	-.1
	2	빈도	560	721	619	270	71	71	172	97	68	3
		수정된 잔차	**2.1**	-2.2	-2.0	1.1	1.5	.0	1.9	.1	.2	.1
	3	빈도	573	788	455	272	63	88	173	118	74	3
		수정된 잔차	**3.3**	1.8	-10.1	1.6	.4	**2.6**	**2.3**	**2.8**	1.2	.1
전체	빈도		1974	2885	2476	962	231	268	573	364	251	11

군집 1의 집단은 최우선 행정 추진분야에 대하여 주민소득 증대의 응답비율이 높고 군집 2의 집단은 환경보존의 응답 비율이 높으며 군집 3의 집단은 환경보존, 치안유지, 교육개선, 행정서비스 개선의 응답 비율이 높다.

5. 결론

본 논문에서는 2003년 조사된 경상남도 사회지표조사 자료에 대하여 k-평균 클러스터링 적용에 대하여 연구하였다. k-평균 클러스터링 적용 시, 군집의 수를 3~9개로 설정하여 각 모형을 탐색 한 후 군집의 속성이 가장 잘 파악되는 군집을 선택하여 심층분석을 실시하였다. k-평균 클러스터링의 탐색 결과 3군집으로 k-평균 클러스터링을 실시했을 때 군집의 속성을 가장 잘 파악할 수 있었으며 카이제곱 검정으로 각 군집에 대한 환경관련문항과의 속성의 차이를 알아 볼 수 있었다. 향후 다양한 데이터에 대하여 k-평균 클러스터링 적용 시, 군집의 수를 다양하게 설정하여 모형을 생성하고 생성된 모형을 비교하여 가장 적절한 군집의 수를 결정하여 결정된 군집의 수에 의한 k-평균 클러스터링 모형 생성 및 분석으로 더욱더 유용한 정보를 추출할 수 있다.

참고문헌

1. Chu, Roddick and Pan(2002). “An Incremental Multi-Centroid, Multi-Run Sampling Scheme for k-medoids-based Algorithms-Extended Report.” *in Second International Conference on Knowledge Discovery and Data Mining,* p553-562.
2. Ester, M., Sander, J. and Kriegel, H. (1999). "A density-based algorithm for discovering clusters in large spatial databases with noise.", *in Second International Conference on Knowledge Discovery and Data Mining*, p226-231.
3. F.Zaki, A.Abd El-Fattah, Y.Enab and S. El-Konyaly(1988). "An ensemble average classifier for pattern recognition machines." Pattern Recognition, vol. 21, no.4, 372-332.
4. Huang, Z. (1997). “Clustering Large Data Sets with Mixed Numeric and Categorical Values.” *In Proceedings of The First Pacific-Asia Conference on Knowledge Discovery and Data Mining,* p21-34
5. Kaufman, L. and Rousseeuw, P.J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons.
6. MacQueen, J. (1967). “Some methods for classification and analysis of multivariate observations.” in *5th Berkeley Symp. Math. statist, Prob*. 1, p281-297.
7. Ng, R. and Han, J. (1994). “Efficient and effective clustering method for spatial data mining.” in *Very Large Data Bases (VLDB'94)*. p144-155.