Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 3, pp. 557~566

Environmental Survey Data Modeling Using K-means Clustering Techniques

Hee-Chang Park[1] · Kwang-Hyun Cho[2]

Abstract

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 analyze 2002 Gyeongnam social indicator survey data using k-means clustering techniques for environmental information. We can use these outputs given by k-means clustering for environmental preservation and environmental improvement.

Keywords : clustering, data mining, environmental information, k-means clustering, modeling

1. 서론

국가에서는 깨끗한 대기질 확보, 수질개선과 양질의 물 공급, 자원 순환형 폐기물 관리 등 환경정책 목표 달성을 위해 환경정보화를 추진하고, 열린 환경 행정 및 대국민 서비스 확대를 위한 정보인프라 구축과 종합적·체계적 환경관리를 위한 환경정보화 기반 마련함으로써 궁극적으로는 생명존중 및 지속가능한 녹색국가 구현을 위한 정보화를 추진목표로 삼고 있다(환경부(2003a)). 이러한 환경정보화는 환경정책 목표 달성을 지원하고, 아울러 환경정보의 활발한 대내외 유통을 통하여 국민의 알권리를 충족하는 동시에 궁극적으로는 환경을 보전하고 개선하는 데 기여하게 된다.

환경데이터에 대해 그 동안 환경관련분야에서는 자료의 수집과 기초적인 분석방법, 다변량 분석방법 등에 중점을 두고 연구가 활발히 이루어지고 있다(이상훈(1995), 이용우(1998), 정상용 등(1998), 최성우와 송형도(2000), 문상기와 우남칠(2001), 환경부(2001), 환경부(2002a), 환경부(2002b), 환경부(2002c), 환경부(2002d), 김정태 등(2003),

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

환경부(2003b), 환경부(2003c) 등). 그럼에도 불구하고 데이터의 양이 기하급수적으로 증가하고 있는 오늘날 방대한 양의 데이터베이스(database : DB)에 내재되어 있는 유용한 정보를 탐색하여 의미 있는 지식을 발견하기 위한 연구의 필요성이 대두되고 있으며, 이를 위한 도구가 데이터마이닝(data mining)이다.

본 논문에서는 데이터들을 k개의 군집으로 임의로 분할을 하여 군집의 평균을 대표값으로 분할해 나가는 방법인 k-평균 클러스터링 기법을 이용하여 환경관련 사회지표 조사데이터에 대한 모형화를 시도하고자 한다. k-평균 클러스터링은 MacQueen(1967)에 의해 처음 소개되었으며, 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 셋에서는 효과적이지 못하다. 그래서 이들은 많은 양의 데이터를 취급하기 위해 CLARA(Clustering LARge Applications) 알고리즘을 제안하였다. Ng와 Han(1994) 그리고 Ester 등(1996)은 CLARA를 더욱 향상시킨 CLARANS(Clustering Large Applications based on RANdomized Search)를 제안하였고, Huang(1998)은 k-means가 연속형 데이터에 대해 한정된 단점을 보완한 연속형과 범주형의 혼합된 데이터에 대한 k-prototypes 알고리즘을 제시하였으며, 동시에 범주형 데이터에 대해서 k-modes 알고리즘을 제시하였다. Chu 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 환경 의식자료에 대하여 내재되어 있는 정보를 추출하기 위하여 k-평균 클러스터링의 모형화 방안에 대하여 연구하고 2002년 조사된 경상남도 사회지표 조사 자료를 바탕으로 k-평균 클러스터링 기법을 적용하고자 한다. 논문의 2절에서는 클러스터링에 대하여 기술하고 3절에서는 경상남도 사회 지표조사 자료의 환경관련 설문에 대하여 k-평균 클러스터링의 모형화 방안을 기술한다. 4절에서는 k-평균 클러스터링을 이용한 자료 분석 결과를 기술한 후, 5절에서 결론을 맺는다.

2. 클러스터링

클러스터링은 다양한 특성을 가진 수많은 데이터를 비슷한 성질의 데이터끼리 묶어주는 데이터마이닝의 한 기법으로서 군집의 수 혹은 군집의 구조에 대한 가정이 없으며, 오직 데이터들 사이의 상사성 또는 비상사성(거리)에 의하여 군집을 형성하고, 형성된 군집의 특성을 파악하여 군집들 사이의 관계를 분석하는 기법이다. 따라서 분류된 군집들은 상호 배타적이어서 한 군집에 속한 개체들은 서로 유사한 성질을 가지고 있으며 이들은 다른 군집에 속한 개체들과 상이한 성질을 가지고 있다. 다음은 기본적인 클러스터링 과정이다.

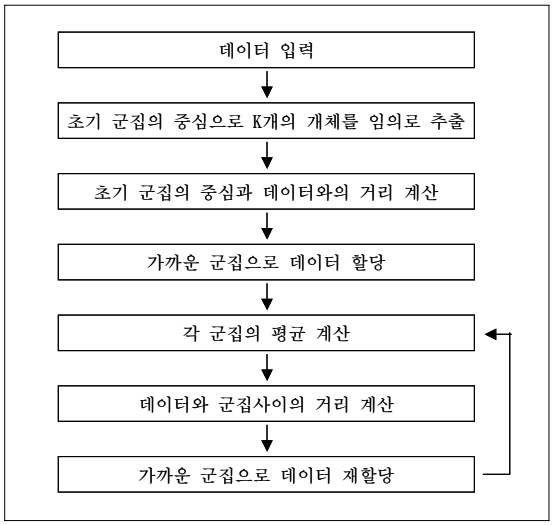
단계 1] N개의 데이터 개체에 대해 p개의 변수를 관찰하여 크기 $(N\times p)$인 자료행렬을 구한다.
단계 2] N개의 데이터 개체 사이의 크기 $(N\times N)$인 거리행렬을 구한다.
단계 3] 거리 행렬로 군집화 방법을 통해서 군집들을 형성한다.
단계 4] 각 군집의 성격이나 상호관계를 분석한다.

클러스트링을 하기 위해서는 먼저 데이터 개체들의 자료정리와 다음에 기술하는 기준 척도를 위한 행렬들을 계산한다. 클러스터링을 수행하기 위한 자료를 정리한 다음에는 묶여지는 각 데이터들 간의 상사성 혹은 비상사성의 정도를 측정하는 기준척도가 필요하다. 두 개체 사이의 거리의 종류에는 유클리드 거리, 유클리드제곱 거리, Mahalanobis 거리, Minkowski 거리 등이 있으나 일반적으로 다음과 같이 정의되는 유클리드 거리를 많이 사용한다.

$$d_{ij}=\sqrt{(X_i-X_j)'(X_i-X_j)} \tag{2.1}$$

기준 척도를 정한 후에는 실제로 대상들을 군집화를 해나가야 한다. 군집화 방법에는 여러 가지가 있으나 크게 계층적 군집화 방법과 비계층적 군집화 방법으로 나누어진다. 계층적 군집화 방법에는 최단 연결법(Single Linkage Method), 최장 연결법(Complete Linkage Method), 평균 연결법(Average Linkage Method), 중심 연결법(Centroid Linkage Method), 중위수 연결법(Median Linkage Method), 그리고 Ward의 방법 등이 있으며, 이는 상사성이 비슷한 군집끼리 묶어 나가는 방법이다. 비계층적 군집화 방법은 통상적으로 미리 규정된 판정기준을 최적화시키도록 시도하고 있고, k-평균 클러스터링 알고리즘과 같이 대부분 연구자에 의해 군집의 개수가 미리 결정되어 있다. 일반적으로 이 방법은 군집의 초기값을 규정하여 결정된 초기군집에 각 개체들을 할당하거나 군집의 일부 개체들 또는 전체를 기준에 따라 최적분리에 이를 때까지 해당하는 규칙에 재 할당 한다. 만약 초기에 부적절한 병합 또는 분리가 일어났을 때 회복될 수 없다는 단점을 최적분리방법은 개체의 재 할당을 통해 극복할 수 있다.

비 계층적 군집화 방법에서 가장 일반적으로 사용되는 k-평균 클러스터링은 데이터들을 k개의 군집으로 임의로 분할을 하여 군집의 평균을 대표값으로 분할해 나가는 방법으로 데이터들을 유사성을 바탕으로 재배치를 하는 방법이다. 알고리즘은 <그림 1> 과 같다.



<그림 1> k-평균 클러스터링 수행 단계

3. k-평균 클러스터링 모형화

본 절에서는 2002년 경상남도 사회 지표조사 자료의 환경관련 설문에 대하여 k-평균 클러스터링의 모형화 방안에 대하여 기술하고자 한다. k-평균 클러스터링 모형화 과정은 <그림 2>와 같으며, 이를 위해 SPSS의 Clementine 10.0을 사용하였다.

<그림 2> k-평균 클러스터링 모형화

1) 자료 구축

k-평균 클러스터링의 모형화에 적용할 자료를 구축하는 과정이다. 자료는 구축은 <표 1>과 같다.

<표 1> 자료 구축

1. 자료 수집	⊙ 2002년 경상남도 사회 지표조사 자료
2. 자료 선정	⊙ 환경관련문항, 인구통계학 속성 문항, 집단구분 문항
3. 자료 정제	⊙ 무응답 등의 결측치 제거

2) 변수 선정

k-평균 클러스터링에 사용할 변수를 선정하고 속성분석에 사용할 변수를 선정한다. 분석 문항 중 주관적 사회계층, 연령, 학력 문항에 대하여 k-평균 클러스터링을 실행하고 환경관련 문항을 각 군집에 대한 속성을 파악하기 위한 변수로 사용한다.

3) k-평균 클러스터링

군집의 특성이 명확하게 파악되는 군집을 도출하기 위하여 3에서 9개의 군집으로 클러스터링을 실행한다.

4) 군집 비교 및 군집 결정

3~9개의 군집으로 k-평균 클러스터링을 실시한 후 군집의 특성이 명확하게 구분되는 군집이 3군집으로 군집-1의 집단은 다른 집단들에 비하여 상대적으로 주관적 사회계층이 낮고 연령이 높으며 학력이 낮은 집단으로 분류되고 군집-2는 다른 집단들에 비하여 상대적으로 주관적 사회 계층이 높고 연령이 낮으면 학력이 높은 집단으로 분류되고 군집-3의 집단은 다른 집단들에 비하여 상대적으로 보통의 성향을 가지는 집단으로 분류되어 3군집의 클러스터링 결과를 분석한다.

5) 군집별 속성 분석

각 군집에 대한 환경 관련 문항에 대하여 차이가 있는지 분석하기 위하여 교차표에 의한 카이제곱 검정이나 분산분석을 실시한다.

4. 자료 분석

자료 분석을 위해 2002년 경남사회지표조사DB로부터 환경관련 문항과 인구통계학 속성 관련 문항, 집단 구분 문항을 추출하여 DB를 구축하였으며, 이 문항들을 환경관련 문항 부문, 분석 문항 부문으로 분류하였다. 설문 문항은 다음과 같다.

<표 2> 경남사회지표조사 환경관련 문항

순번	문항	문항 보기
1	f24. 가장 쾌적한 환경	1.풍부한 녹색 공간 2.맑고 깨끗한 물 3.넓은 공지 4.맑은 공기 5. 기타
2	f25. 수돗물의 음용수 적정 여부	1.적당하다 2.적당하지 않다 3.상수도 시설이 없어 모르겠다 4.상수도 시설이 없지만, 적당하다고 생각한다 5.상수도 시설이 없으며, 적당하지 않다고 생각한다.
3	f26. 수돗물의 음용수 대책	1.상수도 보호구역 확대 지정 2.상수도 시설의 대폭적 개선 3.상수도 환경감시원 제도의 확대 4.기타
4	f27. 환경오염의 주체	1.기업체 2.일반소비자 및 관광행락객 3.농어민 4.모두 5.모르겠음
5	f28. 쓰레기 분리수거 참여 정도	1.잘 참여하고 있다 2.호응하지만 참여정도는 낮다 3.우리 지역은 쓰레기 분리 수거제를 실시하지 않고 있다
6	f29. 녹색제품 구입 여부	1.항상 사려고 노력한다 2.가능하면 사려고 노력한다 3.그런 제품을 사 본 적은 있지만, 굳이 사려고 하지 않는다 4.있는 것은 알지만 사본 적이 없다 5.그런 제품이 있는 것도 모른다

<표 3> 경남사회지표조사 분석 문항

순번	문항	문항 보기
1	g30. 지역사회 전반적 평가	1.아주 살기 좋은 곳이다 2.비교적 살기 좋은 곳이다 3.그저 그런 곳이다 4.비교적 살기 나쁜 곳이다 5.아주 살기 나쁜 곳이다
2	g36. 주관적 사회계층	1.상류층 2.중상류층 3.중류층 4.중하류층 5.하류층
3	j1. 연령	( ) 세
4	j2. 성별	1. 남 2. 여
5	j3. 학력	1. 무학 2. 초졸 3. 중졸 4. 고졸 5. 전문대학재학 6. 전문대졸 7. 대학재학 8. 대졸 9. 대학원이상
6	j6. 직업	1. 의회의원, 고위임직원 및 관리자 2. 전문가 3. 기술공 및 준전문가 4. 사무종사자 5. 서비스 종사자 6. 판매 종사자 7. 농업, 임원 및 어업 숙련 종사자 8. 기능원 및 관련 기능 종사자 9. 장치, 기계 조작 및 조립 종사자 10. 단순노무 종사자 11. 군인 12. 가정주부 13. 학생 14. 무직 15. 기타
7	k1. 조사지역	1. 농촌지역 2. 어촌지역 3. 상가지역 4. 주거지역 5.공업지역 6. 기타지역
8	do. 시 군계	1. 시 2. 군

클러스터링에서는 분석 문항 중 주관적 사회계층, 연령, 학력의 3개의 문항에 대하여 k-평균 클러스터링을 실행하여 군집의 특성이 명확하게 파악되는 k개의 군집으로 나누고 이 군집들의 특성을 파악하여 각 군집별로 환경관련 문항의 응답에 대하여 차이가 있는지를 분석하였다. 군집의 개수를 3에서 9개의 군집으로 클러스터링을 실시한 결과 3개의 군집으로 나누었을 때 군집의 특성이 명확하게 구분되었다. 각 군집의 특성은 <표 2>와 같다.

<표 4> 3군집 클러스터링

군집 / 항목	군집-1	군집-2	군집-3
주관적 사회계층	4.104	3.351	3.641
연령	62.237	32.132	36.514
학력	2.112	7.219	3.855
군집 레코드 수	2810	2410	4657

<표 4>에서 보는 바와 같이 군집-1의 집단은 다른 집단들에 비하여 상대적으로 주관적 사회계층이 낮고 연령이 높으며 학력이 낮은 집단으로 분류되고 군집-2는 다른 집단들에 비하여 상대적으로 주관적 사회 계층이 높고 연령이 낮으면 학력이 높은 집단으로 분류되고 군집-3의 집단은 다른 집단들에 비하여 상대적으로 보통의 성향을 가지는 집단으로 분류되었다. 각 군집에 대하여 환경 관련 문항의 응답의 차이가 있는지 교차표에 의한 카이제곱 검정을 실시하였다. 검정결과 통계적으로 유의한 차이가 있는 문항의 분석 결과만 기술하였다. 세부 내용은 다음과 같다.

<표 5> 각 군집과 쾌적한 자연관경의 조건과의 교차표

군집 \ 쾌적한 자연환경			쾌적한 자연환경의 조건				
			풍부한 녹색공간	맑고 깨끗한 물	넓은 공지	맑은 공기	기타
군집	1	빈도	580	1383	101	690	56
		수정된 잔차	-16.4	**12.4**	.3	**3.2**	**2.0**
	2	빈도	1049	755	67	520	19
		수정된 잔차	**12.7**	-9.4	-2.2	-1.2	-3.6
	3	빈도	1626	1764	179	1005	82
		수정된 잔차	3.9	-3.1	**1.7**	-1.9	1.3
전체	빈도		3255	3902	347	2215	157

<표 5>에서 군집 1의 집단은 쾌적한 자연환경의 조건으로 맑고 깨끗한 물과 맑은 공기, 기타의 응답 비율이 가장 높으며 군집 2의 집단은 풍부한 녹색공간의 응답률이 가장 높으며 군집 3의 집단은 넓은 공지의 응답률이 가장 높은 것으로 나타났다.

<표 6> 각 군집과 수돗물 음용수 적정 여부와의 교차표

군집 \ 수돗물 음용수			수돗물의 음용수 적정여부				
			적당	비적당	상수도 시설 없어 모름	시설 없지만 적당	시설 없지만 비적당
군집	1	빈도	941	1298	211	229	129
		수정된 잔차	**14.3**	-22.9	**10.6**	**10.0**	**3.6**
	2	빈도	397	1849	42	63	58
		수정된 잔차	-9.7	**15.2**	-6.8	-5.7	-3.4
	3	빈도	1010	3150	157	177	162
		수정된 잔차	-4.6	7.6	-3.7	-4.2	-.3
전체	빈도		2348	6297	410	469	349

<표 6>에서 군집 1의 집단은 수돗물 음용수 적정여부에 대하여 적당, 상수도 시설 없어 모름, 시설 없지만 적당, 시설 없지만 비적당의 응답 비율이 높으며 군집 2의 집단은 비적당의 응답비율이 가장 높다.

<표 7> 각 군집과 수돗물의 음용수 대책과의 교차표

군집 \ 수돗물 음용수			수돗물의 음용수 이용대책			
			상수원 확대지정	상수도 시설개선	감시제도 확대	기타
군집	1	빈도	776	1516	441	77
		수정된 잔차	-1.4	**3.0**	-2.8	1.4
	2	빈도	697	1244	415	54
		수정된 잔차	.4	.0	-.3	-.6
	3	빈도	1353	2333	863	106
		수정된 잔차	**.9**	-2.7	**2.8**	-.8
전체	빈도		2826	5093	1719	237

<표 7>에서 군집 1의 집단은 수돗물의 음용수 대책에 대하여 상수도 시설개선의

응답비율이 가장 높았고 군집 3의 집단은 상수원 확대지정, 감시제도 확대의 응답비율이 가장 높다.

<표 8> 각 군집과 환경오염의 해결 주체와의 교차표

군집 \ 환경오염의 주체			환경오염의 주체				
			기업체	일반소비자	농어민	모두	모름
군집	1	빈도	521	380	59	1614	236
		수정된 잔차	**1.9**	**4.5**	**4.1**	-11.8	**15.4**
	2	빈도	391	232	16	1745	26
		수정된 잔차	-1.7	-2.9	-3.3	**7.3**	-7.9
	3	빈도	803	499	58	3190	107
		수정된 잔차	-.3	-1.6	-.8	4.4	-7.1
전체	빈도		1715	1111	133	6549	369

<표 8>에서 군집 1의 집단은 환경오염의 주체에 대하여 기업체, 일반소비자, 농어민, 모름의 응답비율이 가장 높으며 군집 2의 집단은 모두의 응답비율이 가장 높다.

<표 9> 각 군집과 쓰레기 분리수거 참여 정도와의 교차표

군집 \ 쓰레기 분리수거			쓰레기 분리수거의 참여정도		
			잘 참여	보통참여	비분리 수거지역
군집	1	빈도	1815	678	317
		수정된 잔차	-5.2	-2.1	**14.1**
	2	빈도	1712	634	64
		수정된 잔차	**3.1**	.9	-7.9
	3	빈도	3235	1216	206
		수정된 잔차	2.0	**1.1**	-6.0
전체	빈도		6762	2528	587

<표 9>에서 군집 1의 집단은 쓰레기 분리수거의 참여정도에 대하여 비분리 수거지역의 응답 비율이 가장 높으며 군집 2의 집단은 잘 참여의 응답 비율이 높으며 군집 3의 집단은 보통참여의 응답 비율이 가장 높다.

<표 10> 각 군집과 녹색제품 구입여부와의 교차표

군집 \ 녹색제품의 구입			녹색제품의 구입여부				
			항상 구매	가능구매	구매 집착 없음	구매 경험 없음	모름
군집	1	빈도	177	491	381	586	1174
		수정된 잔차	-5.3	-19.5	-3.9	-2.3	**32.0**
	2	빈도	241	1036	391	507	235
		수정된 잔차	**2.6**	**13.3**	.6	-1.9	-15.6
	3	빈도	441	1634	794	1121	666
		수정된 잔차	**2.6**	6.2	**3.1**	**3.7**	-15.5
전체	빈도		859	3161	1566	2214	2075

<표 10>에서 군집 1의 집단은 녹색제품의 구입여부에 대하여 모름의 응답비율이 가장 높으며 군집 2의 집단은 항상 구매와 가능구매의 응답 비율이 가장 높으며 군집 3의 응답 집단은 항상 구매, 구매집착 없음, 구매 경험 없음의 응답 비율이 가장 높다.

5. 결론

본 논문에서는 환경 의식자료에 대하여 데이터마이닝의 기법 중 k-평균 클러스터링의 모형화 방안에 대하여 연구하고 2002년 조사된 경상남도 사회지표 조사의 자료에 대하여 k-평균 클러스터링 기법을 적용하여 모형을 구축하고 구축된 모형을 분석을 분석하였다. 3~9개의 k-평균 클러스터링을 실시하여 비교, 분석한 결과 3개의 군집으로 나누는 것이 구분 문항에 대해 군집의 특성이 가장 명확하게 구분되었다. 군집-1의 집단은 다른 집단들에 비하여 상대적으로 주관적 사회계층이 낮고 연령이 높으며 학력이 낮은 집단으로 분류되고 군집-2는 다른 집단들에 비하여 상대적으로 주관적 사회 계층이 높고 연령이 낮으면 학력이 높은 집단으로 분류되고 군집-3의 집단은 다른 집단들에 비하여 상대적으로 보통의 성향을 가지는 집단으로 분류되었다. 또한 군집별 속성을 분석하기 위해 카이제곱 검정과 Scheffe 검정법을 사용하여 군집별 속성을 파악하였다. 그 결과, 군집-1의 집단은 쓰레기 분리수거를 잘 시행하고 있지 않은 반면 군집-2의 집단은 잘 시행하고 있는 것으로 나타났으며 녹색제품 구입 여부에 대해서는 군집-1의 집단은 구매하지 않는 것으로 나타났으나 군집-2의 집단은 가능한 한 구매하는 것으로 나타났다. 향후 이 정보를 바탕으로 다양한 속성들 간의 분석을 실시하여 환경개선대책 수립과 환경 정책 결정에 필요한 의사결정 지원 등 효율적인 환경행정의 수행과 환경 정책 수립에 기여할 수 있을 것이다.

참고문헌

1. 김정태, 정진도, 김광석(2003), 여름철 충청남도 서북부 지역에서의 대기오염물질 농도 분포특성에 관한 연구, 대한환경공학회 2003 춘계학술발표회 논문집, 1326-1328
2. 문상기, 우남칠(2001), 통계분석을 이용한 지하수위 변동 특성 분류, 한국지하수토양환경학회 01 추계학술발표회논문집, 2001권, 155-159
3. 이상훈(1995), 수질자료의 추세분석을 위한 비모수적 통계검정에 관한 연구, 환경영향평가, 제4권 제2호, 93-103
4. 이용우(1998), 폐기물 배출량의 지역간 차이에 관한 분석, 대한지리학회 33권 2호, 209-224
5. 정상용, 강동환, 심병완(1998), 부산지역 지하수의 수질오염 특성, 한국지하수토양환경학회 98 공동심포지엄 및 추계학술발표회 논문집, 1998권, 86-92
6. 최성우, 송형도(2000), 다변량 통계분석법을 이용한 대구지역 부유분진의 오염원 기여도 추정, 한국환경위생학회지, 제26권 제4호, 1-8

7. 환경부(2001), 전국폐기물통계조사.
8. 환경부(2002a), 전국폐기물발생현황.
9. 환경부(2002b), 상수도통계.
10. 환경부(2002c), 하수도통계.
11. 환경부(2002d), 오수·분뇨 및 축산폐수처리 통계.
12. 환경부(2003a), 2004년도 환경정보화촉진시행계획.
13. 환경부(2003b), 환경통계연감.
14. 환경부(2003c), 대기환경연보.
15. Chu S., Roddick, J.F., Pan, J.(2002). An incremental multi-centroid, multi-run sampling scheme for k-medoids-based algorithms-extended report. *Data Mining III - Proc. Third International Conference on Data Mining Methods and Databases*. Bologna, Italy, 553-562.
16. Ester, M., Kriegel, H., Sander, J., Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise.* The Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 226-231.
17. Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values, *Data Mining and Knowledge Discovery,* 2, 283-304.
18. Kaufman, L. and Rousseeuw, P.J. (1990), *Finding groups in data: an introduction to cluster analysis,* New York, John Wiley & Sons.
19. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, The *5th Berkeley Symp. Math. statist, Prob.* 1, 281-297.
20. Ng, R. and Han, J. (1994). Efficient and effective clustering method for spatial data mining. *Very Large Data Bases (VLDB'94).* 144-155.
21. Zaki, F., El-Fattah A., Enab Y., and El-Konyaly S.(1988). An ensemble average classifier for pattern recognition machines, *Pattern Recognition*, 21(4), 372-332.

[2005년 5월 접수, 2005년 6월 채택]