Proceedings of Joint Conference of Korean Data And Information Science Society and The Korean Data Analysis Society, April 28 - 29, 2006, pp. 241-251

Industrial Waste Database Analysis Using Data Mining

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

Abstract

Data mining is the method to find useful information for large amounts of data in database It is used to find hidden knowledge by massive data, unexpectedly pattern, relation to new rule. The methods of data mining are decision tree, association rules, clustering, neural network and so on.

We analyze industrial waste database using data mining technique. We use k-means algorithm for clustering and C5.0 algorithm for decision tree and Apriori algorithm for association rule. We can use these analysis outputs for environmental preservation and environmental improvement.

Keywords : data mining, association rule, k-means clustering, decision tree.

1. 서론

전국 폐기물 발생 및 처리현황을 조사하는 목적은 전국 생활·사업장폐기물 발생량 및 처리현황을 조사하여 행정구역별, 폐기물종류별 발생량과 처리방법별 처리량을 파악하기 위해서이다. 또한 연도별 폐기물 발생량 및 처리방법의 변화추이를 분석하여 폐기물정책 수립의 기초 자료로 활용하기 위함이다. 2002년 1월 1일부터 12월 31일까지 전국 16개 시·도 및 234개 시·군·구를 대상으로 환경부에서 조사한 2002년도 전국 폐기물 발생 및 처리현황 보고서(환경부와 국립환경연구원, 2002)를 살펴보면 폐기물 발생현황(폐기물 발생현황 추이, 폐기물 종류별 발생현황 분석, 지역별 폐기물 발생량 분석, 폐기물 성상 변화추이 분석), 폐기물 처리현황(폐기물 처리현황 추이, 폐기물 종류별 처리방법 변화 추이, 폐기물 처리주체별 처리현황), 폐기물 처리 관련 시설·장비 현황(처리시설 현황, 폐기물처리업체 현황), 그리고 생활폐기물 관리인원·장비 및 예산현황(지방자치단체의 생활폐기물 관리인원 및 장비, 자치단체 생활폐기물 관리예산) 등에 대해 도표를 통하여 상당히 유익하게 작성되어 있다. 그러나 통계분석을 위해 사용한 방법은 대부분이 일차적인 분석방법에 지나지 않는다. 일차적인 분석만을 하게 되면 조사하는 데 드는 비용을 감안할 때 효과를 충분

[1]Graduate Student, Department of Statistics, Changwon National University, Changwon, Kyungnam, 641-773, Korea. E-mail : cho1023@changwon.ac.kr

[2]First Author : Professor, Department of Statistics, Changwon National University, Changwon, Kyungnam, 641-773, Korea. E-mail : hcpark@changwon.ac.kr

히 얻었다고는 할 수 없을 것이다. 다시 말하면 데이터 내에 잠재되어 있는 더 많은 정보를 추출할 수 있음에도 불구하고 이를 활용하지 못함으로써 기회비용이 커진다고 할 수 있을 것이다. 이에 본 논문에서는 폐기물 데이터베이스에 내재되어 있는 정보를 파악하기 위해 데이터마이닝(data mining) 기법을 이용하여 데이터 분석을 하고자 한다.

데이터마이닝이란 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정으로, 대용량(massive)의 관측 가능한 데이터를 기반으로 숨겨진 지식, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 바탕으로 의사결정 등을 위한 정보로 활용하고자 하는 것이다. 데이터 마이닝이 적용되는 과정은 탐색(exploration)을 통해 평균, 이상치, 결측치 등을 발견하고 변형(modification)으로 자료를 변환하며, 모형화(modeling)와 모델평가(assessment)의 단계를 거치게 된다.

본 논문에서는 데이터마이닝 알고리즘 중 가장 많이 이용되는 연관규칙, 의사결정나무 기법, K-평균군집방법을 이용하여 환경 데이터에 대한 분석을 하고자 한다. 이를 위해 먼저 환경부에서 조사한 2002년도 전국 폐기물 발생현황에 대한 데이터베이스에 있는 경상남도에 소속한 시군의 산업폐기물 데이터와 경상남도 통계 데이터베이스(경상남도, 2002)에 있는 지역 여건 데이터를 추출한 후, 이들을 통합하여 데이터베이스화하였다. 또한 데이터마이닝 분석을 위한 소프트웨어로는 SPSS사의 Clementine을 이용하였다. 본 논문의 2절에서는 K-평균 군집기법을 이용하여 산업폐기물 자료를 분석한 결과를 기술한다. 3절에서는 의사결정나무기법을 이용하여 분석한 결과를 제시한다. 4절에서는 연관성 규칙을 이용한 자료 분석 결과를 기술한 후, 5절에서 결론을 맺는다.

2. K-평균 군집 기법을 이용한 산업폐기물 자료 분석

데이터마이닝은 방대하고 다양한 형태의 데이터로부터 의사결정에 유용한 정보를 발견하려는 일련의 데이터 분석 및 모형선정 기법이다. 데이터마이닝의 기법 중에서 군집화는 다양한 특성을 지닌 관찰대상을 유사성을 바탕으로 동질적인 집단으로 분류하는데 쓰이는 기법으로 이를 크게 나누면 분할 군집법과 계층적 군집법이 있다. 그 중에서 분할 군집법은 데이터들을 임의의 부분집합으로 분할을 한 후 데이터들을 유사한 그룹으로 재배치하는 군집방법이다. 분할 군집법의 종류에는 본 논문에서 고려한 k-평균 알고리즘과 k-medoids 알고리즘, k-prototypes 알고리즘, k-modes 알고리즘 등이 있다. 이들 중에서 k-평균 알고리즘은 MacQueen(1967)에 의해 처음 소개된 것으로, 데이터들을 k개의 군집으로 임의로 분할을 하여 군집의 무게중심(평균)을 대표값으로 하여 분할해 나가는 방법으로, 데이터들을 유사성을 바탕으로 재배치를 하는 방법이다.

군집분석에서 군집간의 유사성 측정은 거리로써 나타낸다. 서로 다른 개체 사이의 거리 $d_{ij} = d(X_i, X_j)$를 구하는 방법에는 유클리디안(Euclidean) 거리, 유클리디안 제곱거리, 마할라노비스(Mahalanobis) 거리, 그리고 민코우스키(Minkowski) 거리 등이 있으며, 본 논문에서는 다음과 같이 정의되는 유클리디안 제곱거리를 이용하고자 한다.

$$d_{ij} = \sum_{k=1}^{p} (t_{ik} - t_{jk})^2 \quad (2.1)$$

본 절에서는 사용한 데이터는 환경부에서 조사한 2002년도 전국 폐기물 발생현황에 대한 데이터베이스를 이용하였으며, 이 데이터베이스에서 경상남도 시군 레코드(시군의수 : 20개)의 산업 폐기물 관련 항목들만 선정하여 산업 생활계 폐기물 발생량과 산업 배출 시설계 폐기물 발생량 항목에 대하여 3개의 군집으로 k-평균 군집화를 실시하였다. 여기서 산업 생활계 폐기물은 사무실, 식당, 기숙사 등에서 발생되는 폐기물이고 산업 배출시설계 폐기물은 배출시설의 설치·운영과 관련하여 배출되는 폐기물이다. 산업 생활계 폐기물 관련 항목에 대한 3-평균 군집화의 결과는 <표 1>과 같다.

<표 1> 산업 생활계 폐기물관련 항목에 대한 3-평균 군집화(단위: 톤/일)

군집 / 항목	군집-1	군집-2	군집-3
자치처리단체의 매립량	2.081	18.1	1.0
자치처리단체의 소각량	1.263	3.6	4.7
자치처리단체의 재활용량	1.031	3.333	5.0
처리업체처리 매립량	0.119	0.3	5.7
처리업체처리의 소각량	0.331	0	6.6
처리업체처리의 재활용량	6.387	5.067	8.4
자가처리의 소각량	0.119	2.987	2.6
자가처리의 재활용량	0.262	3.733	0.2
군집 레코드 수	16	3	1
시군명	창원시, 진주시, 진해시, 사천시, 밀양시, 거제시, 양산시, 의령군, 함안군, 고성군, 남해군, 하동군, 산청군, 함양군, 거창군, 합천군	마산시, 통영시, 창녕군	김해시

<표 1>에서 보는 바와 같이 군집-1은 전체적으로 산업 생활계 폐기물에 대한 처리량이 다른 집단들보다 상대적으로 낮은 집단으로 나타났다. 이들 군집에 속하는 시군은 창원시, 진주시, 진해시, 사천시, 밀양시, 거제시, 양산시, 의령군, 함안군, 고성군, 남해군, 하동군, 산청군, 함양군, 거창군, 합천군 등이다. 군집-2는 다른 집단에 비해 상대적으로 자치처리단체의 매립량과 자가처리의 소각량 및 자가처리의 재활용량이 높으며 처리업체의 소각량이 낮은 집단으로 마산시, 통영시, 창녕군이 군집-2에 속한다. 군집-3은 자치처리단체의 매립량 및 자가처리의 재활용량이 가장 낮으며 자치처리단체의 소각량 및 재활용량, 처리업체의 매립량 및 소각량, 재활용량이 높은 집단으로 김해시가 군집-3에 속한다.

각 군집에 대하여 경상남도에서 조사한 경상남도 통계 데이터베이스의 산업 관련 지역 여건 항목에 대한 통계량(평균)은 <표 2>와 같다.

산업 배출시설계 폐기물관련 문항에 대한 3-평균 군집화의 결과는 <표 3>과 같다.

<표 2> 산업 생활계 폐기물에 대한 군집별 지역여건 통계량

항목 \ 군집	군집-1	군집-2	군집-3
전체 행정구역 면적(km2)	559.8	336.5	463.26
전체 행정구역 인구수(명)	131467.6	212892.7	390292
농업 및 임업사업체수(개)	9.7	4.7	4
어업 사업체수(개)	1.1	1.7	0
광업 사업체수(개)	3.5	1.7	6
제조업 사업체 수(개)	900.6	1204.7	4759
전기, 가스, 수도 사업체 수(개)	4.9	7.3	13
건설업 사업체 수(개)	236.1	292.3	523
도매, 소매업 사업체 수(개)	2344.7	4431.7	5596
숙박 및 음식점업 사업체 수(개)	2008.9	3536.7	4769
운수업 사업체 수(개)	623.3	1444.7	1916
통신업 사업체 수(개)	26.4	4	36
사업서비스업 사업체 수(개)	150.4	263.7	316

<표 3> 산업 배출시설계 폐기물관련 항목에 대한 3-평균 군집화(단위: 톤/일)

항목 \ 군집	군집-1	군집-2	군집-3
자치처리단체의 매립량	1.76	4.914	46.7
자치처리단체의 소각량	0	0.129	9.8
자치처리단체의 재활용량	0	0.286	0.0
처리업체처리의 매립량	181.04	26.264	93.6
처리업체처리 소각량	8.2	5.336	23.4
처리업체처리의 재활용량	695.26	70.372	279.699
자가처리의 매립량	202.039	0.0	0.0
자가처리의 소각량	17.64	1.664	13.0
자가처리의 재활용량	344.88	2.093	77.0
군집 레코드 수	5	14	1
시군명	창원시, 거제시, 함안군, 고성군, 하동군	마산시, 진주시, 진해시, 통영시, 사천시, 밀양시, 양산군, 의령군, 창녕군, 남해군, 산청군, 함양군, 거창군, 합천군	김해시

<표 3>에서 보는 바와 같이 군집-1은 전체적으로 산업 배출시설계 폐기물에 대하여 처리업체의 처리량과 자가처리의 처리량이 높으며 자치처리단체의 처리량은 낮은 집단으로 나타났으며, 창원시, 거제시, 함안군, 고성군, 하동군이 군집-1에 속한다. 군집-2는 전체적으로 산업 배출시설계 폐기물에 대한 처리량이 다른 집단들보다 상대적으로 낮은 집단으로 나타났으며 마산시, 진주시, 진해시, 통영시, 사천시, 밀양시, 양산군, 의령군, 창녕군, 남해군, 산청군, 함양군, 거창군, 합천군이 군집-2에 속한다. 군집-3은 다든 군집에 비하여 자치처리단체의 매립량 및 소각량이 높은 집단으로 나타났으며 김해시가 군집-3에 속한다.

각 군집에 대하여 경상남도에서 조사한 경상남도 통계 데이터베이스의 산업 관련 지역 여건 항목에 대한 통계량(평균)은 다음과 같다.

<표 4> 산업 배출시설계 폐기물에 대한 군집별 지역여건 통계량

항목 \ 군집	군집-1	군집-2	군집-3
전체 행정구역 면적(km2)	461.1	553.6	463.26
전체 행정구역 인구수(명)	176400.4	132868.4	390292
농업 및 임업사업체수(개)	12.2	8.4	4
어업 사업체수(개)	2.0	0.9	0
광업 사업체수(개)	3.0	3.3	6
제조업 사업체 수(개)	1196.6	860.1	4759
전기, 가스, 수도 사업체 수(개)	5.2	5.3	13
건설업 사업체 수(개)	320.6	218.0	523
도매, 소매업 사업체 수(개)	2877.4	2601.6	5596
숙박 및 음식점업 사업체 수(개)	2549.6	2143.1	4769
운수업 사업체 수(개)	701.6	771.4	1916
통신업 사업체 수(개)	32.3	27.2	36
사업서비스업 사업체 수(개)	255.2	137.2	316

3. 의사결정나무기법을 이용한 산업폐기물 자료 분석

의사결정나무는 의사결정규칙(decision rule)을 나무구조로 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법이다. 분석과정이 나무구조에 의하여 표현되므로 분류 또는 예측을 목적으로 하는 방법인 회귀분석(regression analysis), 신경망모형, 판별분석에 비해 연구자는 분석과정을 쉽게 이해하고 설명할 수 있다. 이러한 의사결정나무분석을 수행하기 위해 다양한 분리기준, 정지규칙, 가지치기 방법이 제안되어 있으며, 이들을 어떻게 결합하느냐에 따라서 서로 다른 의사결정나무 형성방법이 만들어진다. 의사결정나무의 진행과정은 다음과 같다.

의사결정나무의 생성 ----> 가지치기 ---> 타당성 평가 ---> 해석 및 예측

의사결정나무분석의 대표적인 알고리즘으로는 Kass(1980)가 제안한 CHAID(Chi-squared Automatic Interaction Detection), Breiman 등(1984)이 제안한 CART(classification and regression trees), 그리고 Quinlan(1993)이 제안한 C5.0 등이 있다.

CART 알고리즘은 의사결정나무분석을 형성하는데 있어서 가장 보편적인 알고리즘이라고 할 수 있으며, 기계-학습(machine-learning) 실험의 시초가 되고 있다. 모형의 형성은 훈련자료집합(training data set)을 가지고 한다. CART는 목표변수가 범주형인 경우에는 지니지수(Gini index)를 적용하고, 목표변수가 연속형인 경우에는 분산의 감소량을 이용하여 이지분리(binary split)를 수행하는 알고리즘이다. 지니지수는 불순도(impurity)를 측정하는 하나의 지수로 다음과 같이 정의된다.

$$G = \sum_{j=1}^{c} \sum_{i \neq j} P(i)P(j) \tag{3.1}$$

여기서 $P(i)$는 각 마디에서 한 개체가 목표변수의 i번째 범주에 속할 확률이다. 따라서 지니지수는 임의의 한 개체가 목표변수의 i번째 범주로부터 추출되었고, 그 개체를 목표변수의 j 번째 범주에 속한다고 오분류(misclassification)할 확률인 $P(i)P(j)$의 합으로 표현된다. 여기서 c는 목표변수의 범주의 수를 말한다. 지니지수의 값이 작을수록 한 범주가 지배적으로 많게 되어 그 노드의 순수도는 높아지므로 지니지수는 불순도를 측정한다고 할 수 있다. 이를 근거로 의사결정나무를 형성하면 순수도를 최대로 하는 하위노드를 갖게 하는 분리가 타당하게 되며, 불순도를 가장 많이 줄여 줄 수 있는 설명변수가 영향을 가장 많이 주는 변수가 된다.

C5.0 알고리즘은 기계학습 분야의 효력 있는 ID3(induction of decision tree) 알고리즘의 확장된 기법이다. C5.0는 ID3에서 다룰 수 없었던 값, 즉 연속형 변수를 다룰 수 있으며, 의사결정나무에서 가지치기(pruning)을 할 수 있는 등 ID3를 보완한 알고리즘이다. C5.0에서는 엔트로피(entropy)를 불확실성의 측도로 사용하게 되는데, 이 개념은 정보이론(Information thoery)으로부터 생성되었다. 엔트로피는 노드의 순수도를 측정하는 도구로 다음과 같이 정의된다.

$$E = -\sum_{i=1}^{c} p_i \log_2(p_i) \tag{3.2}$$

엔트로피는 전달할 수 있는 정보와 이에 대응하는 확률을 곱해서 더한 것이므로 기대정보(expected information)라고 할 수 있다. 노드가 순수하다면 표현하는 범주의 수는 감소하므로 엔트로피의 값이 줄어들게 되므로 엔트로피는 노드의 순수도를 측정하는 도구라고 할 수 있다. 따라서 의사결정나무는 엔트로피가 감소하는 방향으로 자라게 되며, 엔트로피를 가장 많이 줄여주는 변수를 기준으로 형성된다. C5.0이 앞에서 기술한 CART와의 다른 점은 CART는 이진(binary)분리를 하는 반면에 C5.0은 여러 개의 가지를 가진다는 것이다.

CHIAD는 카이제곱-검정(이산형 목표변수) 또는 F-검정(연속형 목표변수)을 이용하여 다지 분리(multiway split)를 수행하는 알고리즘이다. 이 알고리즘은 두 변수간의 통계적 관계를 찾는 것으로 AID(automatic interaction detection system)에 기원을 두고 있다. CHIAD는 각 설명변수의 범주들이 자료를 반응변수의 각 범주들로 구분하는 판별력의 크기에 따라 설명변수의 범주들을 이용하여 나무구조를 만드는 분석방법으로 전체 자료를 둘 이상의 하위노드(child node)로 반복적으로 분할한다. 이 과정에서 설명변수의 범주의 쌍에 대한 반응변수의 유의한 차이가 없으면 설명변수의 범주들을 병합하며, 유의적이지 않은 쌍들이 없을 때까지 과정을 계속한다. 각 설명변수에 대한 최고의 분할을 찾고, 모든 설명변수에 대한 유의성을 조사하여 가장 유의적인 설명변수를 선택한다. 선택된 설명변수의 범주들의 그룹을 사용해 자료를 상호 배반인 부분집합으로 분할하며 각 부분집합에서 정지규칙중의 하나가 만족될 때까지 이 과정을 독립적으로 순환, 반복한다. CART와 다른 점은 CHAID는 데이터를 overfitting 하기 전에 나무 형성을 멈춘다는 것이다.

본 절에서 사용한 데이터는 2절에서 사용한 데이터와 경상남도에서 조사한 경상남도 통계 데이터베이스의 지역 여건 항목을 추출한 후, 이들을 통합하여 데이터베이스화하였다. 이 항목들 중에

서 산업 생활계 폐기물 발생량과 산업 배출시설계 발생량 항목에 대하여 평균이하와 평균초과의 이분형으로 변환하여 C5.0 알고리즘을 이용하여 의사결정나무 분석을 실시하였다. 산업 생활계 폐기물에 대한 의사결정나무 기법을 적용하기 위한 데이터 속성은 다음과 같다.

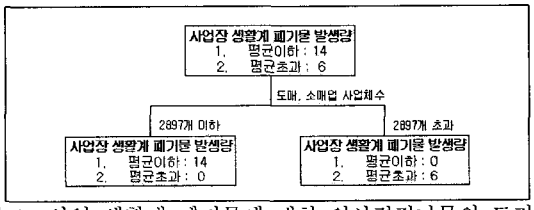
◉ 목표변수

1. 산업 생활계 폐기물 발생량 (1) 평균 이하 (2) 평균 초과

◉ 예측변수

1. 시군 구분
2. 전체 행정구역 면적
3. 전체 행적구역 인구수
4. 농업 및 임업사업체수
5. 어업 사업체수
6. 광업 사업체수
7. 제조업 사업체 수
8. 전기, 가스, 수도 사업체 수
9. 건설업 사업체 수
10. 도매, 소매업 사업체 수
11. 숙박 및 음식점업 사업체 수
12. 운수업 사업체 수
13. 통신업 사업체 수
14. 사업서비스업 사업체 수

산업 생활계 폐기물에 대한 의사결정나무의 트리 구조는 <그림 1>과 같다. <그림 1>에서 보는 바와 같이 산업 생활계 폐기물 발생량에 대하여 예측변수 15항목 중 도매, 소매업 사업체 수 항목만이 목표변수에 영향을 주며 도매, 소매업 사업체의 수가 2897개 이하이면 산업 생활계 폐기물 발생량이 평균 이하인 시군이고, 도매, 소매업 사업체수가 2897개 초과이면 산업 생활계 폐기물 발생량이 평균 초과인 시군으로 구분할 수 있다.

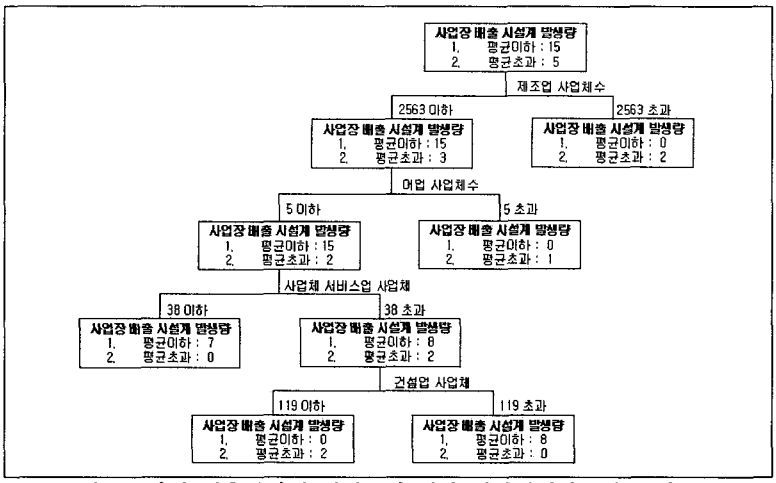


<그림 1> 산업 생활계 폐기물에 대한 의사결정나무의 트리 구조

산업 배출시설계 폐기물에 대한 의사결정나무 기법을 적용하기 위한 데이터 속성은 목표변수가 산업 배출시설계 폐기물 발생량 항목이며, 예측변수는 산업 생활계 폐기물 발생량에 대한 의사결정나무의 예측변수와 동일하다. 산업 배출시설계 폐기물에 대한 의사결정나무의 트리 구조는 <그림 2>와 같다.

<그림 2>에서 보는 바와 같이 산업 배출 시설계 폐기물 발생량에 대하여 예측변수 15항목 중 제조업 사업체수, 어업사업체수, 사업 서비스업 사업체수, 건설업 사업체 수가 목표변수에 영향을 주고 있다. 제조업 사업체 수가 2536개를 초과하면 산업 배출 시설계 폐기물 발생량이 평균 초과인 시군이고, 제조업 사업체수가 2536개 이하이고 어업 사업체수가 5개 초과이면 산업 배출 시설계 폐기물 발생량이 평균 초과인 시군이다. 또한 제조업 사업체수가 2536개 이하이고 어업 사업체수가 5개 이하이고 사업 서비스업 사업체가 38개 초과이고 건설업 사업체수가 119이하이면 산업

배출 시설계 폐기물 발생량이 평균 초과인 시군이다. 반면에 제조업 사업체 수가 2536개 이하이고 어업 사업체수가 5개 이하이며 사업 서비스업 사업체수가 38개 이하이면 산업 배출 시설계 폐기물 발생량이 평균 이하인 시군으로 구분할 수 있다. 제조업 사업체 수가 2536개 이하이고 어업 사업체수가 5개 이하이며, 사업 서비스업 사업체수가 38개 초과이고 건설업 사업체 수가 119 미만이면 산업 배출 시설계 폐기물 발생량이 평균 이하인 시군으로 구분할 수 있다.



<그림 2> 산업 배출시설계 폐기물에 대한 의사결정나무의 트리 구조

4. 연관성 규칙을 이용한 산업폐기물 자료 분석

Agrawal 등(1993)에 의해 처음 소개된 연관성 규칙은 각 항목간의 연관성을 반영하는 규칙으로서 둘 또는 그 이상의 항목간의 연관성 규칙의 평가기준을 기반으로 하여 의미 있는 규칙을 찾아내는 데이터마이닝 기법 중의 하나이다. 연관성 규칙의 평가기준에는 지지도(support), 신뢰도(confidence), 향상도(lift) 등이 있다. 지지도는 항목 집합 X와 항목 집합 Y가 동시에 발생한 비율을 의미하며, 식 (4.1)과 같이 정의된다.

$$S_{(X \Rightarrow Y)} = P(X \cap Y) \tag{4.1}$$

신뢰도는 항목집합 X가 포함된 비율 중 항목 집합 X와 Y가 동시에 포함된 비율을 의미하며, 식 (4.2)와 같이 정의된다.

$$C_{(X \Rightarrow Y)} = P(Y \mid X) = \frac{P(X \cap Y)}{P(X)} \tag{4.2}$$

향상도는 실제발생 확률을 각 항목집합의 발생이 독립적일 경우 그 거래가 동시에 발생할 예상 기대확률로 나눈 것을 의미하며, 식 (4.3)과 같이 정의된다.

$$L_{(X \Rightarrow Y)} = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)} \quad (4.3)$$

향상도의 값이 1보다 크면 두 항목이 동시에 발생한 확률이 예상확률보다 더 크므로 향상도의 값이 1이상인 경우에 의미 있는 관련성 규칙이라고 할 수 있다. 그 값은 많은 관련성 규칙의 조합 중에서 의미 있는 관련성규칙을 발견하기 위한 기준으로서 많이 사용된다.

연관성 규칙은 항목 집합간의 지지도와 신뢰도를 계산하여, 미리 분석자에 의해서 정해진 최소 지지도 및 최소 신뢰도를 모두 만족하는 두 항목 집합의 규칙을 강한 연관 규칙을 가진 것으로 판단한다. 연관성 규칙을 해석할 때 신뢰도의 값이 크면 좋지만 신뢰도가 크다고 모두 최선의 연관성 규칙은 아니다. 두 항목의 집합의 기본적인 지지도가 어느 정도 수준 이상일 경우만 고려해야 한다. 또한 신뢰도와 지지도는 자주 발생되는 항목 집합에 대해서는 연관성 때문이 아니라 우연하게 높게 나올 수도 있으므로 향상도를 잘 관찰할 필요가 있다.

본 절에서 사용한 데이터는 3절에서 사용한 데이터와 동일한 것으로, 산업 폐기물 관련 항목의 연관성 규칙과 산업 폐기물관련 항목과 지역 여건 항목들 간의 연관성 규칙을 유도하였다. 산업 폐기물 관련 항목의 연관성 규칙은 <표 5>와 같다. <표 5>에서 알 수 있는 바와 같이 산업 생활계 폐기물 발생량은 산업 생활계 폐기물 재활용량과 산업 생활계 폐기물 매립량이 관련성이 있는 것으로 나타났다. 또한 산업 생활계 폐기물 발생량은 산업 생활계 폐기물 재활용량과 산업 생활계 폐기물 소각량, 산업 생활계 폐기물 매립량과 산업 생활계 폐기물 소각량, 산업 생활계 폐기물 재활용량과 산업 생활계 폐기물 매립량이 관련성이 있는 것으로 나타났다. 산업 배출시설계 폐기물 발생량은 산업 배출시설계 폐기물 매립량, 산업 배출시설계 폐기물 매립량과 산업 배출시설계 폐

<표 5> 산업 폐기물 관련 항목의 연관성 규칙

규칙	지지도	신뢰도	후항값	전항값1	전항값2
1	30	100	산업 생활계 폐기물 발생량 : 평균 초과	산업 생활계 폐기물 재활용량 : 평균 초과	
2	15	100	산업 생활계 폐기물 발생량 : 평균 초과	산업 생활계 폐기물 매립량 : 평균 초과	
3	25	100	산업 생활계 폐기물 발생량 : 평균 초과	산업 생활계 폐기물 재활용량 : 평균 초과	산업 생활계 폐기물 소각량 : 평균 초과
4	10	100	산업 생활계 폐기물 발생량 : 평균 초과	산업 생활계 폐기물 매립량 : 평균 초과	산업 생활계 폐기물 소각량 : 평균 초과
5	15	100	산업 생활계 폐기물 발생량 : 평균 초과	산업 생활계 폐기물 재활용량 : 평균 초과	산업 생활계 폐기물 매립량 : 평균 초과
6	20	100	산업 배출시설계 폐기물 발생량 : 평균 초과	산업 배출시설계 폐기물 매립량 : 평균 초과	
7	20	100	산업 배출시설계 폐기물 발생량 : 평균 초과	산업 배출시설계 폐기물 매립량 : 평균 초과	산업 배출시설계 폐기물 재활용량 : 평균 초과
8	15	100	산업 배출시설계 폐기물 발생량 : 평균 초과	산업 배출시설계 폐기물 재활용량 : 평균 초과	산업 배출시설계 폐기물 소각량 : 평균 초과
9	10	100	산업 배출시설계 폐기물 발생량 : 평균 초과	산업 배출시설계 폐기물 매립량 : 평균 초과	산업 배출시설계 폐기물 소각량 : 평균 초과

기물 재활용량, 산업 배출시설계 폐기물 재활용량과 산업 배출시설계 폐기물 소각량, 그리고 산업 배출시설계 폐기물 매립량과 산업 배출시설계 폐기물 소각량이 관련성이 있는 것으로 나타났다.

산업 폐기물 관련 항목과 지역여건 항목에 대해 연관성 규칙을 생성한 결과는 <표 6>과 같다. <표 6>에서 보는바와 같이 산업 생활계 폐기물 발생량은 제조업 사업체의 수, 사업서비스업 사업체의 수,, 숙박 및 음식점업 사업체의 수와 사업서비스업 사업체의 수, 건설업 사업체의 수와 사업서비스업 사업체의 수, 그리고 행정구역 인구수와 사업서비스업 사업체의 수가 관련성이 있는 것으로 나타났다.

<표 6> 산업 폐기물 관련 항목과 지역여건 항목의 연관성 규칙

규칙	지지도	신뢰도	후항값	전항값1	전항값2
1	25	100	산업 생활계 폐기물 발생량 : 평균 초과	제조업 사업체 : 평균 초과	
2	20	100	산업 생활계 폐기물 발생량 : 평균 초과	사업서비스업 사업체 : 평균 초과	
3	10	100	산업 생활계 폐기물 발생량 : 평균 초과	숙박 및 음식점업 사업체 : 평균 초과	사업서비스업 사업체 : 평균 초과
4	20	100	산업 생활계 폐기물 발생량 : 평균 초과	건설업사업체 : 평균 초과	사업서비스업 사업체 : 평균 초과
5	20	100	산업 생활계 폐기물 발생량 : 평균 초과	행정구역 인구수 : 평균 초과	사업서비스업 사업체 : 평균 초과

5. 결론

본 연구에서는 2002년 환경부에서 발표한 전국 폐기물 발생현황의 경상남도 관련 산업 폐기물에 대하여 데이터 마이닝 기법을 이용하여 분석을 실시하였다. 그 결과 군집화 기법을 이용하여 폐기물 관련 항목에 대하여 시군별로 비슷한 성질을 가지는 군집으로 나누고 각 군집에 있는 집단들의 성향을 파악할 수 있었다. 의사결정나무 기법을 이용하여 폐기물 관련 항목에 대하여 각 항목별로 영향을 주는 항목을 찾아낼 수 있었으며, 각 항목에 대한 속성별로 분류를 할 수 있어서 시도의 폐기물 관련 현황을 쉽게 파악할 수 있었다. 연관성 규칙을 이용하여 각 항목에 대한 관련성 여부를 파악할 수 있어서 폐기물 관련 항목들 간에 유기적인 관계를 파악할 수 있었다. 이러한 데이터마이닝 기법을 이용한 분석 자료는 폐기물 관련 환경개선이나 정책결정 등에 도움을 줄 수 있을 것으로 사료된다.

참고문헌

[1] 경상남도 (2002). *경남통계연보*, 경상남도기획관실.

[2] 환경부, 국립환경연구원 (2003). *2002 전국 폐기물 발생 및 처리현황*.

[3] Agrawal, R., Imielinski, R., and Swanmi, A. (1993). Mining association rules between sets of items in large

databases, *Proceeding of the ACM SIGMOD Conference on Management of Data*, Washington, D.C. (SIGMOD93), p207-216

[4] Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone (1984). *Classification and regression trees*, Wadsworth, Belmont.

[5] Kass, G. V. (1980). *An Exploratory Technique for Investigating Large Quantities of Categorical Data*

[6] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. in *5th Berkeley Symp. Math. statist, Prob.* 1, 281-297.

[7] Quinlan, J. R. (1993). *C4.5 Programs for Machine Learning*. San Mateo, Morgan Kaufmann.