Waste Database Analysis Joined with Local Information Using Decision Tree Techniques

Hee Chang Park[1], Kwang Hyun Cho[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.

The decision tree approach is most useful in classification problems and to divide the search space into rectangular regions. Decision tree algorithms are used extensively for data mining in many domains such as retail target marketing, fraud detection, data reduction and variable screening, category merging, etc. We analyze waste database united with local information using decision tree techniques for environmental information. We can use these decision tree outputs for environmental preservation and improvement.

Keywords : CART, C5.0, decision trees, environmental information

1. 서론

전국 폐기물 발생 및 처리현황을 조사하는 목적은 전국 생활·사업장폐기물 발생량 및 처리현황을 조사하여 폐기물종류별, 행정구역별 발생량과 처리방법별 처리량을 파악하기 위해서이다. 또한 연도별 발생량 및 처리방법의 변화추이를 분석하여 폐기물 정책 수립의 기초 자료로 활용하기 위함이다(환경부, 국립환경연구원 (2003)). 그동안 환경부에서 조사한 전국 폐기물 발생 및 처리현황 보고서를 살펴보면 폐기물 발생현황, 폐기물 처리현황, 폐기물 처리 관련 시설·장비 현황, 그리고 생활폐기물 관리인원·장비 및 예산현황 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 충분히 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)기법 중 하나인 의사결정나무기법(decision

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

2) Graduate Student, Department of Statistics, Changwon National University, Changwon, Gyungnam, 641-773, Korea

tree techniques)을 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. 데이터마이닝 기법으로는 군집분석(cluster analysis), 연결 분석(link analysis), 판별 분석(discrimination analysis), 연관성규칙(association rule), 의사결정나무기법, 신경망모형(neural network) 등의 분석 기법이 있다.

본 논문에서 적용한 의사결정나무(decision tree)는 데이터마이닝에서 사용하는 여러 기법들 중의 하나이며, 의사결정 규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법으로 다른 분석방법에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점이 있다.

그동안의 연구를 살펴보면 의사결정나무분석을 수행하기 위한 다양한 분리기준, 정지규칙, 가지치기 방법들이 제안되었으며, 이들을 어떻게 결합하느냐에 따라서 서로 다른 의사결정나무가 형성된다. 또한 정확하고 빠르게 의사결정나무를 형성하기 위해 다양한 알고리즘이 제안되고 있다. 의사결정나무 알고리즘에는 Hartigan(1975)이 제안한 CHAID(Chi-squared Automatic Interaction detecion), Breiman(1984)이 제안한 CART(Classification and Regression Trees), Quinlan(1993)에 의해 제안된 C4.5, 그리고 Lon와 Shin(1997)이 제안한 QUEST(Quick, Unbiased, Efficient, Statistical Tree) 알고리즘 등이 있다.

본 논문에서는 경상남도 폐기물 관련 자료에 내제되어 있는 의미 있는 정보를 파악하기 위하여 전국 폐기물 발생현황 자료에 의사결정나무기법을 적용하고자 한다. 이를 위해 1999년~2002년 환경부에서 조사한 전국 폐기물 발생현황 데이터베이스에서 경상남도에 소속한 시군의 폐기물 데이터와 1999년~2002년 경상남도에서 발표한 경남통계 연보 데이터베이스의 지역여건 데이터를 통합하여 새로운 데이터베이스를 구축한 후 의사결정나무 기법을 적용한다. 통합된 데이터베이스에 대하여 의사결정나무기법을 적용함으로서 폐기물 관련 자료를 더욱 더 자세하게 분류 및 세분화를 할 수 있다. 본 논문의 2절에서는 의사결정나무기법을 소개한 후, 3절에서 의사결정나무기법을 적용하기 위한 자료 통합과정을 기술한다. 4절에서는 의사결정나무기법을 이용한 자료 분석 결과를 제시한 다음, 5절에서 결론을 맺는다.

2. 의사결정나무기법

의사결정나무는 의사결정규칙을 나무구조형태로 도표화하여 관심의 대상이 되는 집단을 여러 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석기법이다. 의사결정나무(decision tree) 탐색(exploration)과 모형화(modeling)라는 두 가지 특성을 모두 가지고 있다고 할 수 있다. 의사결정나무는 분류 또는 예측의 과정이 나무구조에 의한 추론규칙(induction rule)에 의해서 표현되기 때문에, 신경망이나, 판별분석, 회귀분석 등 다른 방법에 비해서 그 과정을 쉽게 이해하고 설명할 수 있다. 또한 의사결정나무는 신경망과는 달리 규칙(rule)방식으로 표현이 가능하며, 규칙은 SQL과 같은 database access 언어로 표현될 수 있다는 장점을 가지고 있다. 의

사결정나무는 판별분석(discrimination analysis) 또는 회귀분석(regression analysis) 등과 같은 모수적(parametric) 모형을 분석하기 위해서 사전에 이상치(outlier)를 검색하거나 분석에 필요한 변수 또는 모형에 포함되어야 할 교호효과를 찾아내기 위해서 사용될 수도 있고, 그 자체가 분류 또는 예측 모형으로 사용될 수도 있다. 의사결정나무를 형성할 때 목표변수가 범주형인 경우에는 분류나무(classification tree)를 형성한다고 하며, 또 연속형인 경우에는 회귀나무(regression tree)를 형성한다고 한다. 일반적으로 의사결정나무분석의 절차는 다음과 같다.

[단계 1] 의사결정나무의 형성 : 분석목적과 자료의 구조에 따라 적절한 분리기준(split criterion)과 정지규칙(stopping rule)을 고려하여 의사결정나무를 형성한다.

[단계 2] 가지치기 및 축소 : 부적절한 추론규칙(induction rule)을 가지고 있거나 분류오류(classification error)를 크게 할 위험이 있는 가지를 제거한다. 또한 가지들간의 관계를 조절하는 가지축소(shrinking) 과정을 거친다.

[단계 3] 타당성 평가 : 위험도표(risk chart), 이익도표(gains chart), 또는 교차타당성 평가 등을 통하여 의사결정나무를 평가한다.

[단계 4] 모형의 해석 및 예측 : 얻어진 의사결정나무모형을 해석하고 예측모형을 설정한다.

본 논문에서는 여러 가지 의사결정나무 알고리즘 중에서 CART와 C5.0 알고리즘을 이용하여 환경의식자료의 모형화를 시도하고자 한다. CART는 목표변수가 이산형인 경우에는 불순도(impurity)를 측정하는 지니지수(Gini index)를 이용하고, 연속형인 경우에는 분산의 감소량을 이용하여 이지분리(binary split)를 수행하는 알고리즘이다. 지니지수는 각 마디에서의 불순도 또는 다양도(diversity)를 측정하는 것으로 식 (2.1)과 같이 표현된다.

$$G = \sum_{i=1}^{c} P(i)\,(1 - P(i)) \tag{2.1}$$

여기서 c는 목표변수의 범주수이고, $P(i)$는 목표변수에 의해 분할된 c개 부그룹의 비율을 의미한다. 반면에 C5.0은 다지분리를 수행하는 알고리즘으로 엔트로피(entropy)를 불확실성의 측도로 이용하여 예측변수의 기준으로 사용한다. 엔트로피는 식(2.2)와 같이 정의된다.

$$E = \sum_{i=1}^{c} P(i)\,(-\log_2 P(i)) \tag{2.2}$$

3. 자료 통합

3.1 자료 현황

본 논문에서는 환경부에서 발표한 1999년부터 2002년까지의 전국 폐기물 발생 현황에 대한 데이터를 사용한다(환경부 국립환경연구원(1999), 환경부 국립환경연구원(2000), 환경부 국립환경연구원(2001), 환경부 국립환경연구원(2002)). 전국 폐기물 발생 현황에 대한 데이터 구조는 <표 1>과 같다. 또한 경상남도에서 발표한 1999년부터 2002년까지의 경남통계 연보 데이터를 사용한다(경상남도(1999), 경상남도(2000), 경상남도(2001), 경상남도(2002)). 경남통계 연보의 데이터 구조는 <표 2>와 같다.

<표 1> 전국 폐기물 발생 현황 데이터 구조

측정분야	측정 항목
생활폐기물	1. 생활폐기물 - 발생량 - 각 단체별 처리량, 처리방법 별 처리량 2. 사업장 생활계폐기물 - 발생량 - 각 단체별 처리량, 처리방법 별 처리량
사업장배출시설계폐기물	- 발생량 - 각 단체별 처리량, 처리방법 별 처리량
건설폐기물	- 발생량 - 각 단체별 처리량, 처리방법 별 처리량
인원 및 장비 현황	- 각 단체별 인원, 차량, 손수레, 중장비 현황
수집 운반차량	- 각 단체별 압착압축, 압롤, 덤프 현황

<표 2> 경남통계 연보 데이터 구조

측정 분야	연혁, 도지 및 기후, 인구, 노동, 사업체, 농림수산업, 광업・제조업 및 에너지, 전기・가스・수도, 유통・금융・보험 및 기타 서비스, 주택・건설, 교통・관광 및 정보통신, 보건 및 사회보장, 환경, 교육 및 문화, 재정, 소득, 공공행정 및 사법의 총 17개 분야

3.2 자료의 통합

전국 폐기물 발생 현황에서의 경상남도에 소속된 시군에 대한 총 4개의 폐기물 항목과 경남통계 연보에서의 인구 및 사업체 분야에 대한 총 10개의 지역여건 항목을 추출하여 폐기물 발생 항목과 지역 여건 항목으로 통합된 데이터베이스를 구축하였

다. 데이터 융합에 사용한 폐기물 및 지역 여건 항목은 <표 3>과 <표 4>와 같다.

<표 3> 폐기물 항목

순번	항목
1	생활폐기물 발생량
2	사업장 생활계 폐기물 발생량
3	사업장배출시설계 폐기물 발생량
4	건설폐기물 발생량

<표 4> 지역 여건 항목

순번	항목
1	인구수
2	가구수
3	시도구분
4	농업, 임업 사업체수
5	어업 사업체수
6	광업 사업체수
7	제조업 사업체수
8	건설업 사업체수
9	도매 및 소매업 사업체수
10	숙박 및 음식점업 사업체수

3.3 자료의 변환

의사결정나무 분석 시 모형의 결과 해석의 용이함을 고려하여 폐기물 데이터에 대하여 평균을 바탕으로 평균이하와 평균초과의 이분형으로 자료를 변환하였다. 의사결정나무 분석을 위해 <표 5>의 변환된 자료를 바탕으로 모형구축 및 규칙생성을 하였다.

<표 5> 폐기물 데이터 변환

순번	항목	변환
1	생활폐기물 발생량	1. 평균이하 2. 평균초과
2	사업장 생활계 폐기물 발생량	1. 평균이하 2. 평균초과
3	사업장배출시설계 폐기물 발생량	1. 평균이하 2. 평균초과
4	건설폐기물 발생량	1. 평균이하 2. 평균초과

3.4 의사결정나무 적용

의사결정나무 적용 과정은 <그림 1>과 같다.

<그림 1> 의사결정나무 모형화

[단계 1] 자료 구축
의사결정나무 분석에 적용할 자료를 구축한다. 구축된 자료는 <표 6>와 같다.

<표 6> 자료 구축

1. 자료 수집	⊙ 1999년부터 2002년까지의 전국 폐기물 발생 현황 데이터 ⊙ 1999년부터 2002년까지의 경남 통계 연보 데이터
2. 자료 통합	⊙ 폐기물 관련 항목과 지역여건 항목의 자료 통합
3. 자료 정제	⊙ 무응답 등의 결측치 제거
4. 자료 변환	⊙ 폐기물 관현 항목을 이분형으로 변환

[단계 2] 변수 선정
의사결정나무에 사용할 목표변수와 입력변수를 선정한다. 분석에 사용한 변수는 <표 7>과 같다.

<표 7> 목표변수, 입력변수 선정

출처	변수	항목
환경부	목표변수	생활폐기물 발생량, 사업장 생활계 폐기물 발생량, 사업장 배출시설계 폐기물 발생량, 건설 폐기물 발생량
경상남도 통계 DB	입력변수	인구수, 가구수, 시도구분, 농업 및 임업 사업체수, 어업 사업체수, 광업 사업체수, 제조업 사업체수, 건설업 사업체수, 도매 및 소매업 사업체수, 숙박 및 음식점업 사업체수

[단계 3] 알고리즘 선정
의사결정나무 분석 시 이지 분리가 가능한 CART 알고리즘을 사용하여 분석을 실시하였다.

[단계 4] 모형 구축
모형구축 시 훈련자료와 모형 평가자료로 구분하고 각각 2/3, 1/3로 자료를 분할한다. 가지치기를 위하여 가지치기 강도를 60~80으로 설정하고 최소레코드수를 5로 설정하여 각각의 모형을 구축한다. 이때, 가지치기 강도와 최소 레코드수를 낮게 설정하면 모형의 정확도는 증가하나 모형이 복잡해짐으로써 모형의 해석이 어려워질 수 있다.

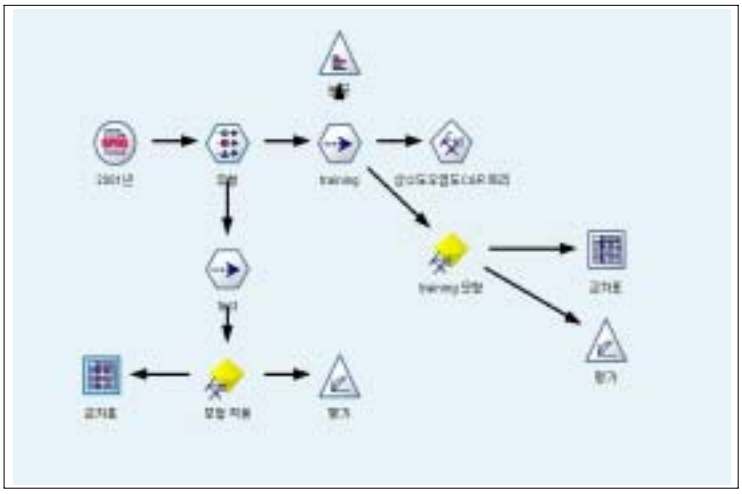
[단계 5] 모형 비교
가지치기 강도 60~80으로 설정한 의사결정나무 모형을 비교한다. 본 논문에서는 가지치기 강도를 75로 설정하였을 때 의미 있는 규칙을 발견할 수 있었다.

[단계 6] 모형 생성
의사결정나무 모형을 생성하고 분석한다.

4. 의사결정나무에 의한 결과 해석

4.1 의사결정나무 모형 생성 스트림

SPSS의 Clementine 10.0을 이용한 의사결정나무 모형 생성 스트림은 <그림 2>와 같다.



<그림 2> 의사결정나무 모형 생성 스트림

4.2 의사결정나무 분석 결과

생활폐기물 발생량, 사업장 생활계폐기물 발생량, 사업장 배출시설계 폐기물 발생량, 건설폐기물 발생량에 대한 의사결정나무 분석 결과는 다음과 같다.

1) 생활폐기물 발생량
생활폐기물 발생량에 대한 결과는 <표 8>과 같다.

<표 8> 생활폐기물 발생량 의사결정나무 분석 결과

규칙	노드	목표변수
1	⊙ 숙박 및 음식점업 사업체(2631.5 이상)	평균↓ : 0% **평균↑ : 100%**
2	⊙ 숙박 및 음식점업 사업체(2631.5 미만) → 농업 및 임업 사업체 (0.5 미만)	평균↓ : 0% **평균↑ : 100%**

숙박 및 음식점업 사업체 수가 2631.5개 이상인 집단은 생활폐기물 발생량에 대하여 평균초과가 100%이다. 또한, 숙박 및 음식점업 사업체 수가 2631.5개 미만이고 농업 및 임업 사업체 수가 0.5개 미만인 집단은 생활폐기물 발생량에 대하여 평균초과가 100%이다.

2) 사업장 생활계폐기물 발생량
사업장 생활계 폐기물 발생량에 대한 결과는 <표 9>과 같다.

<표 9> 사업장 생활계폐기물 발생량 의사결정나무 결과 분석

규칙	노드	목표변수
1	⊙ 도매 및 소매업 사업체(2883.5 이상)	평균↓ : 11.8% **평균↑ : 88.2%**
2	⊙ 도매 및 소매업 사업체(2883.5 이상) → 농업 및 임업 사업체 (25 미만)	평균↓ : 6.25% **평균↑ : 93.75%**
3	⊙ 도매 및 소매업 사업체(2883.5 이상) → 농업 및 임업 사업체 (25 미만) → 제조업 사업체 (3814.5 이상) → 가구수 (102932 이상)	평균↓ : 0% **평균↑ : 100%**

도매 및 소매업 사업체 수가 2883.5개 이상인 집단은 사업장 생활계폐기물 발생량에 대하여 평균초과가 88.2%이다. 도매 및 소매업 사업체 수가 2883.5개 이상이고 농업 및 임업 사업체 수가 25개 미만인 집단은 사업장 생활계폐기물 발생량에 대하여 평균초과가 93.75%이다. 또한, 도매 및 소매업 사업체 수가 2883.5개 이상이고 농업 및 임업 사업체 수가 25개 미만이고 제조업 사업체 수가 9184.5개 이상이면서 가구수가 102932가구 이상이면 사업장 생활계폐기물 발생량에 대하여 평균초과가 100%이다.

3) 사업장 배출시설계 폐기물 발생량
사업장 배출시설계 폐기물 발생량에 대한 결과는 <표 10>와 같다.

<표 10> 사업장 배출시설계 폐기물 발생량 의사결정나무 결과 분석

규칙	노드	목표변수
1	⊙ 제조업 사업체(2438 이상)	평균↓ : 11.1% **평균↑ : 88.9%**

제조업 사업체 수가 2438개 이상인 집단은 사업장 배출시설계 폐기물 발생에 대하여 평균초과가 88.9%이다.

4) 건설 폐기물 발생량

건설폐기물 발생량에 대한 결과는 <표 11>과 같다.

<표 11> 건설폐기물 발생량 의사결정나무 결과 분석

규칙	노드	목표변수
1	⊙ 제조업 사업체(666 이상)	평균↓ : 0% **평균↑ : 100%**
2	⊙ 제조업 사업체(666 미만) → 가구수 (43588.5 이상)	평균↓ : 0% **평균↑ : 100%**
3	⊙ 제조업 사업체(666 미만) → 가구수 (43588.5 미만) → 농업 및 임업 사업체 (8.5 이상 14.5 미만) → 광업사업체(5.5 이상)	평균↓ : 0% **평균↑ : 100%**

제조업 사업체 수가 666개 이상인 집단은 건설 폐기물 발생량에 대하여 평균 초과가 100%이다. 제조업 사업체 수가 666개 미만이고 가구수가 43588.5개 이상인 집단은 건설 폐기물 발생량에 대하여 평균 초과가 100%이다. 또한, 제조업 사업체 수가 666개 미만이고 가구수가 43588.5가구 미만이고 농업 및 임업 사업체 수가 8.5개 이상 14.5개 미만이면서 광업 사업체 수가 5.5개 이상인 집단은 건설 폐기물 발생량에 대하여 평균 초과가 100%이다.

5. 결론

본 논문에서는 경남지역 폐기물 데이터 분석을 위하여 1999년부터 2002년까지의 전국 폐기물 발생 현황 데이터와 1999년부터 2002년까지의 경상남도에서 발표한 경남통계 연보 데이터를 통합하여 새로운 데이터베이스를 구축하고 구축된 데이터베이스에 대하여 의사결정나무 기법을 적용하였다. 통합된 데이터베이스에 대하여 의사결정나무 기법의 모형 생성 및 분석 결과, 각 폐기물 발생량에 대한 현황을 쉽게 파악 할 수 있었고 폐기물 발생량에 따른 지역 여건 항목의 관련 정도 및 세분화 등의 유용한 정보를 추출할 수 있었다. 이러한 의사결정나무 기법을 이용한 지역정보와 통합한 폐기물 데이터 분석 자료는 폐기물 관련 환경개선이나 정책결정 등에 도움을 줄 수 있

을 것으로 사료된다.

참고문헌

1. 경상남도(1999), 경남통계연보, 경상남도기획관실.
2. 경상남도(2000), 경남통계연보, 경상남도기획관실.
3. 경상남도(2001), 경남통계연보, 경상남도기획관실.
4. 경상남도(2002), 경남통계연보, 경상남도기획관실.
5. 환경부 국립환경연구원 (2000), 1999 전국 폐기물 발생 및 처리현황.
6. 환경부 국립환경연구원 (2001), 2000 전국 폐기물 발생 및 처리현황.
7. 환경부 국립환경연구원 (2002), 2001 전국 폐기물 발생 및 처리현황.
8. 환경부 국립환경연구원 (2003), 2002 전국 폐기물 발생 및 처리현황.
9. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and regression trees*, Wadsworth and books.
10. Hartigan, J.A. (1975), *Clustering Algorithms*, New York: John Wiley & Sons, Inc.
11. Loh, W.Y and Shin, Y.S(1997). Split Selection Methods for Classification Tree, *Statistica Sinica*, p815-840
12. Quinlan, J.R. (1993), *C4.5 Programs for Machine Learning,* Morgan Kaufmann Publishers.