Proceedings of the Korean Data And Information Science Society,
October 22-23, 2004, pp. 63-75

Modeling of Environmental Survey by Decision Trees[1)]

Hee-Chang Park[2)], Kwang-Hyun Cho[3)]

Abstract

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 dection, data reduction and variable screening, category merging, etc. We analyze Gyeongnam social indicator survey data using decision tree techniques for environmental information. We can use these decision tree outputs for environmental preservation and improvement.

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

1. 서론

환경정보화의 목표는 최신 정보기술을 이용하여 환경관련 정보를 생산·수집·가공·처리·유통 또는 활용함으로써 생명존중 및 지속가능한 녹색국가 구현을 위한 정보화를 추진하는 데 있다(환경부(2003)). 이러한 환경정보화는 환경정책 목표달성을 지원하고, 아울러 환경정보의 활발한 대내외 유통을 통하여 국민의 알권리를 충족하는 동시에 궁극적으로는 환경을 보전하고 개선하는 데 기여하게 된다. 환경문제를 보다 효과적으로 해결하고 예방하기 위해서는 더욱 광범위한 정보를 적기에 확보하여 환경행정에 활용할 수 있는 체계를 구비하여 환경정책에 필요한 고급적인 정보를 도출하여야 한다.(환경부(2004)). 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

---

1 This work has been performed with the financial assistance of Gyeongnam Regional Environmental Technology Development Center (GRETeC), KOREA

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

3) Graduate Student, Department of Statistics, Changwon National University, Changwon, Kyungnam, 641-773, Korea

(2001), 환경부(2002a), 환경부(2002b), 환경부(2002c), 환경부(2002d), 김정태 등(2003), 환경부(2003a), 환경부(2003b) 등). 그럼에도 불구하고 환경분야에서는 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 논문에서 적용한 의사결정나무(decision tree)는 데이터마이닝에서 사용하는 여러 기법들 중의 하나이며, 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. 대표적인 의사결정나무알고리즘에는 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 Loh와 Shin(1997)이 제안한 QUEST(Quick, Unbiased, Efficient, Statistical Tree) 알고리즘 등이 있다.

본 논문에서는 환경 의식자료에 대하여 데이터 속에 내제 되어있는 정보를 추출하기 위하여 데이터마이닝의 기법 중 의사결정나무의 모형화 방안에 대하여 연구하고 2002년 조사된 경상남도 사회지표 조사 자료를 바탕으로 의사결정나무 기법을 적용하여 모형을 구축하고 구축된 모형을 분석하고자 한다. 논문의 2절에서는 의사결정나무에 대하여 기술하고 3절에서는 경상남도 사회 지표조사 자료의 환경관련 설문에 대하여 의사결정나무의 모형화 방안을 기술한다. 4절에서는 의사결정나무를 이용한 자료분석 결과를 기술한 후, 5절에서 결론을 맺는다.

2. 의사결정나무

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. 의사결정나무는 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. 의사결정나무 모형화

본 절에서는 2002년 경상남도 사회 지표조사 자료의 환경관련 설문에 대하여 의사결정나무의 모형화 방안에 대하여 기술하고자 한다. 이를 위해 2002년 경상남도에서 조사한 도민 생활수준 및 의식조사에 대한 자료를 환경관련 문항과 집단구분 문항들을 추출하여 DB화 하였으며, 분석에 사용한 설문 문항은 다음과 같다.

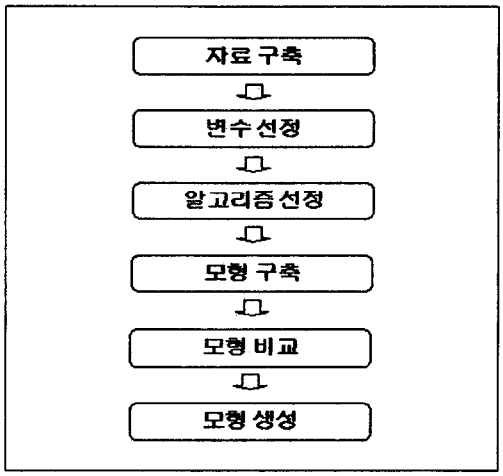
<표 1> 환경관련 문항

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
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 분리수거제를 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 2> 집단구분 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, 기계조작 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

의사결정나무 모형화를 위하여 SPSS의 Clementine 10.0을 사용하였으며, 모형화 과정은 <그림 2>와 같다.



<그림 2> 의사결정나무 모형화

1) 자료 구축

의사결정나무 모형화에 적용할 자료를 구축하는 과정이다. 자료는 구축은 <표 3>과 같다.

<표 3> 자료 구축

1. 자료 수집	■ 2002년 경상남도 사회 지표조사 자료
2. 자료 선정	■ 환경관련문항과 집단구분 문항 선정
3. 자료 정제	■ 무응답 등의 결측치 제거
4. 자료 변환	■ 환경관련문항을 이분형으로 변환

2) 변수 선정

의사결정나무의 모형을 생성하기 위하여 목표변수와 입력변수를 선정한다. 목표변수는 환경관련 문항으로 선정하고 입력변수는 집단구분 문항으로 설정한다.

3) 알고리즘의 선정

의사결정나무 알고리즘 중 2002년 경남 사회지표 조사 자료에 적합한 알고리즘을 선택하기 위하여 다지분리가 가능한 C5.0과 이지분리가 가능한 CART 알고리즘을 사용하여 모형을 구축한다.

4) 모형 구축

C5.0과 CART 알고리즘으로 모형을 구축한다. 모형구축 시 훈련자료와 모형 평가자료로 구분하고 각각 2/3, 1/3로 자료를 분할하고 가지치기를 위하여 가지치기 강도를 75로 설정하고 최소레코드수를 5로 설정한다. 이때, 가지치기 강도와 최소 레코드수를 낮게 설정하면 모형의 정확도는 증가하나 모형이 복잡해짐으로써 모형의 해석이 어려워질 수 있다.

6) 모형비교

C5.0과 CART에 의하여 구축된 의사결정 나무 모형을 비교한다. 다음은 C5.0과 CART의 모형에 대한 예측 정확도와 모형 평가의 예측정확도를 나타난 표이다.

<표 4> C5.0과 CART 모형 예측도 비교

기법 / 항목	C5.0		CART	
	모형 예측정확도	모형평가 예측정확도	모형 예측정확도	모형평가 예측정확도
1. f24_1 가장 쾌적한 환경	60.7%	54.1%	63.2%	59.1%
2. f25 수돗물의 음용수 적정 여부	67.3%	61.9%	66.5%	66.2%
3. f26 수돗물의 음용수 대책	62.2%	50.8%	53.6%	51.1%
4. f27. 환경오염의 주체	65.3%	51.1%	66.0%	66.9%
5. f28. 쓰레기 분리수거 참여 정도	68.9%	56.4%	68.8%	68.2%
6. f29. 녹색제품 구입 여부	67.7%	60.5%	63.5%	62.4%

<표 4>에서 보는 바와 같이 C5.0의 모형과 CART의 모형 예측 정확도가 비슷한 것으로 나타났다. 두 모형 중 어느 것을 사용해도 무방하나 모형 결정에 있어서 모형 예측정확도와 더불어 모형의 평가 또한 중요하게 생각해야 한다. C5.0은 구축된 트리 모형을 검정 자료에 적용한 결과 모형평가 예측정확도가 많이 떨어지는 것을 알 수 있다. 이것은 모형이 안정화되지 못했다는 증거이다. 하지만, CART는 모형평가 예측정확도가 모형 예측정확도와 거의 비슷한 것으로 보아 모형이 안정화되었음을 의미한다. 이와 같이, C5.0 모형과 CART 모형 중 모형이 안정화 되어있는 CART의 모형을 사용하는 것이 바람직하고 본 절에서는 CART를 사용한 의사결정나무 모형에 대하여 분석을 하였다.

7) 모형 생성

모형비교에 의하여 CART의 의사결정 나무 모형이 C5.0의 나무 모형보다 더 좋은 것으로 나타났다. 각각의 환경관련 문항에 대하여 CART 알고리즘을 이용하여 모형을 생성한다.

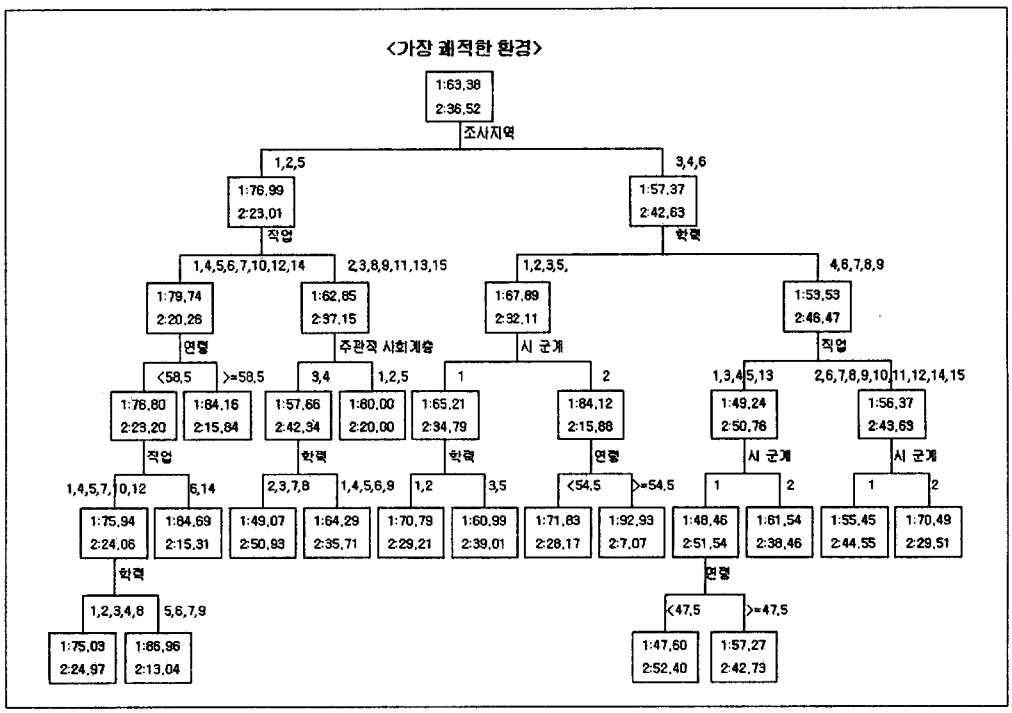
4. 자료 분석

의사결정나무에서는 환경 관련 문항들을 목표변수로 그 외 변수들을 입력 변수로 선택하여 C5.0 기법과 CART 기법으로 분석을 실시하고자 한다. 분석 결과 목표변수들의 보기 항목들이 많아서 모형이 복잡해지고 결과 해석이 어려워서 목표변수인 환경관련 문항들을 다음과 같이 변환하였다.

<표 5> 목표변수의 변환값

순번	문항	문항 보기	변경내용
1	f24_1. 가장 쾌적한 환경	1.넓은 공지 또는 풍부한 녹색공간 2.맑고 깨끗한 물과 공기	1 • 3 → 1, 2 • 4→ 2, 5 → 결측치
2	f25. 수돗물의 음용수 적정 여부	1. 적당하다 2. 그렇지 않다.	1 → 1, 2 • 3 • 4 • 5 → 2
3	f26. 수돗물의 음용수 대책	1.상수원 보호구역 및 환경감시원 확대 2.상수도 시설의 대폭적 개선	1 • 3 → 1 , 2 → 2, 4 → 결측치
4	f27. 환경오염의 주체	1모두 2.기업체, 일반소비자 및 관광 행락객, 농어민 각각	4 → 1, 1 • 2 • 3 → 2, 5 → 결측치
5	f28. 쓰레기 분리수거 참여 정도	1. 잘 참여하고 있다 2. 참여도 낮음	1 → 1 , 2 → 2, 3 → 결측치
6	f29. 녹색제품 구입 여부	1.구매하려고 노력한다. 2. 구매하려지 않는다.	1 • 2 → 1, 3 • 4 • 5→ 2

각 문항에 대한 의사결정나무 모형의 분석은 다음과 같다. 먼저 가장 쾌적한 환경에 대한 의사결정나무 모형은 <그림 3>과 같다.

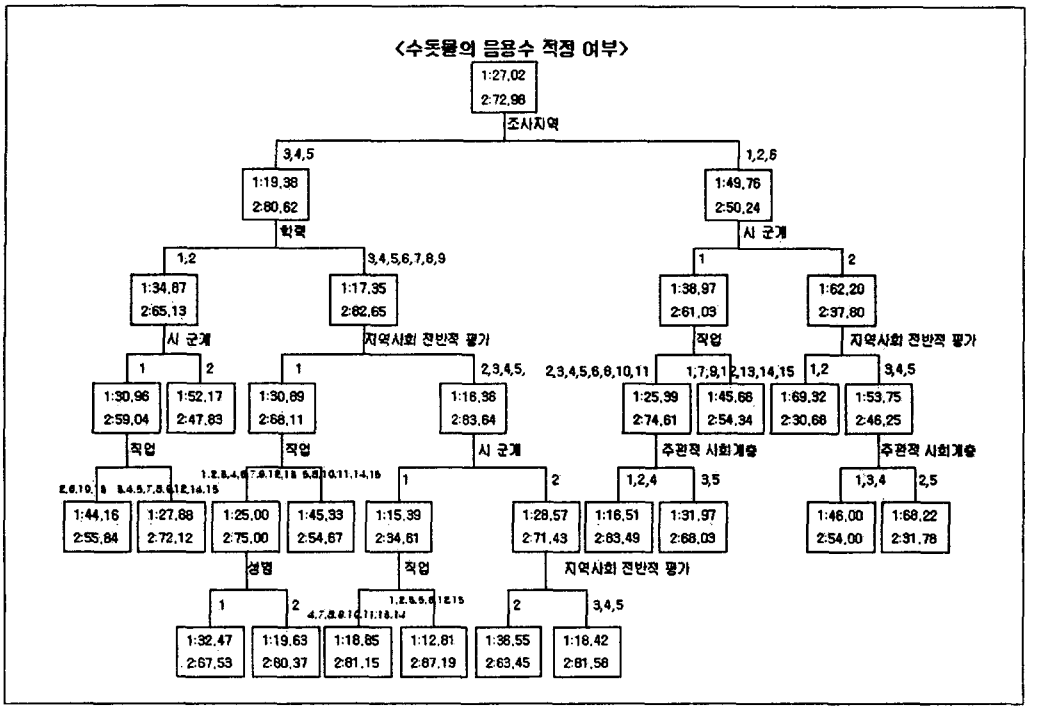


<그림 3> 가장 쾌적한 환경에 대한 의사결정나무 모형

<그림 3>에서 보는 바와 같이 최 상위 노드에서 가장 쾌적한 환경에 대하여 넓은 공지 또는 풍부한 녹색공간이 63.38%, 맑고 깨끗한 물과 공기가 36.62%로 전체적으로 63.38%의 응답자들이 가장 쾌적한 환경으로 넓은 공지 또는 풍부한 녹색공간으로 생각하고 있다. 세부적으로 살펴보면 거주지역이 농촌지역, 어촌지역, 공업지역이고 직업

이 의회의원, 고위임직원 및 관리자, 사무종사자, 서비스 종사자, 농업, 임업 및 어업 숙련 종사자, 단순노무종사자, 가정주부이고 연령이 58세 이하이고 학력이 전문대재학, 전문대졸, 대학재학, 대학원 이상인 응답집단은 가장 쾌적한 환경에 대하여 넓은 공지 또는 풍부한 녹색공간 86.96%로 응답과 거주 지역이 상가지역, 주거지역, 기타지역이고 학력이 중졸이하, 전문대재학이고 거주지가 군이고 연령이 55세 이상인 응답집단은 가장 쾌적한 환경에 대하여 넓은 공지 또는 풍부한 녹색공간 92.93%로 가장 쾌적한 환경에 대하여 넓은 공지 또는 풍부한 녹색공간에 아주 많은 비중을 두는 것으로 나타났다. 반면에 거주지역이 상가지역, 주거지역, 기타지역이고 학력이 고졸, 전문대졸 이상이고 직업이 의회의원, 고위임직원 및 관리자, 기술공 및 준전문가, 사무종사자, 서비스 종사자, 학생이고 거주지가 시이고 연령이 47세 이하인 응답 집단은 가장 쾌적한 환경에 대하여 맑고 깨끗한 물과 공기 52.40%로 가장 쾌적한 환경에 대하여 맑고 깨끗한 물과 공기의 응답률이 증가한 것을 알 수 있다.

수돗물의 음용수 적정여부에 대한 의사결정나무 모형은 <그림 4>와 같다.

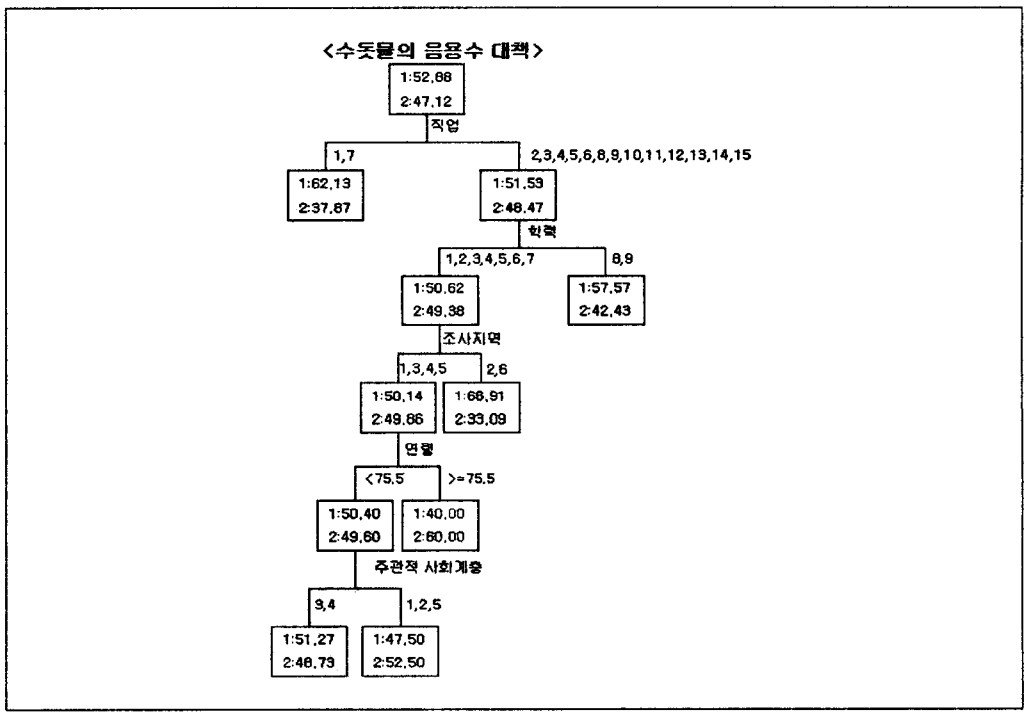


<그림 4> 수돗물의 음용수 적정여부에 대한 의사결정나무 모형

<그림 4>에서 보는 바와 같이 최 상위 노드에서 최 상위 노드에서 수돗물의 음용수 적정여부에 대하여 적당하다가 27.02%, 그렇지 않다가 72.98%로 전체적으로 72.98%의 응답자들이 수돗물의 음용수 적정여부에 대하여 전체적으로 적당하지 않다는 시각을 가지고 있다. 세부적으로 살펴보면, 거주지역이 상가지역, 주거지역, 공업지역이고 학력이 중졸이상이고 지역사회의 전반적 평가가 좋음 이하이고 거주지가 시이고 직업이 의회의원, 고위임직원 및 관리자, 전문가, 기술공 및 준전문가, 서비스 종사

자, 판매종사자, 가정주부, 기타인 응답 집단은 수돗물 음용수 적정 여부에 대하여 적당하지 않다 87.19%로 수돗물 음용수 적정 여부에 대하여 더 부정적인 시각을 가지고 있다. 반면에 거주지역이 농촌지역, 어촌지역, 기타이고 거주지가 군이고 지역사회의 전반적 평가가 좋음 이상인 응답자 집단은 수돗물 음용수 적정 여부에 대하여 적당하다 69.32%의 응답과 거주지역이 농촌지역, 어촌지역, 기타이고 거주지가 군이고 지역사회의 전반적 평가가 보통이하이고 주관적 사회계층이 중상류층, 하류층인 응답자 집단은 수돗물 음용수 적정 여부에 대하여 적당하다 68.22%로 수돗물 음용수 적정 여부에 대하여 긍정적인을 가지고 있다.

수돗물의 음용수 대책에 대한 의사결정나무 모형은 <그림 5>와 같다.

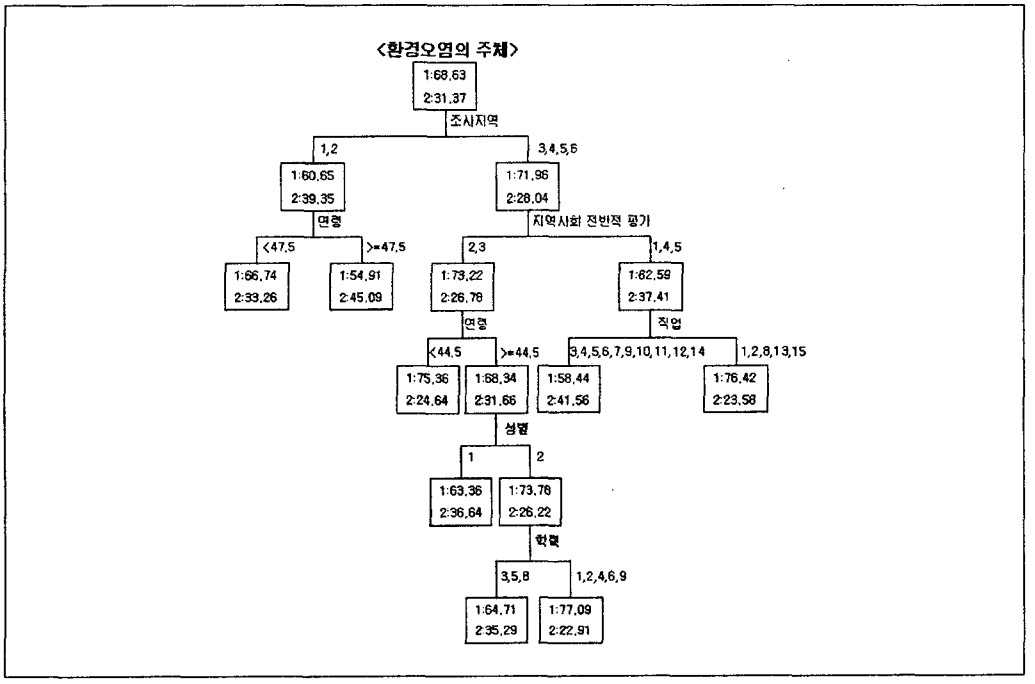


<그림 5> 수돗물의 음용수 대책에 대한 의사결정나무 모형

<그림 5>에서 살펴보면 최 상위 노드에서 수돗물의 음용수 대책에 대하여 1.상수원 보호구역 및 환경감시원 확대가 52.88% 2. 상수도 시설의 대폭적 개선이 47.12%로 전체적으로 수돗물의 음용수 대책에 대하여 비슷한 응답 결과를 보이고 있다. 세부적으로 살펴보면, 직업이 의회의원, 고위 임직원 및 관리자, 농업, 임업 및 어업 숙련 종사 제외이고 학력이 대졸이상 제외이고 거주지역이 어촌지역, 기타지역인 응답집단은 수돗물의 음용수 대책에 대하여 상수원 보호구역 및 환경감시원 확대 66.91%로 수돗물의 음용수 대책에 대하여 상수원 보호구역 및 환경감시원 확대에 더 비중을 주는 것으로 나타났다. 반면에 직업이 의회의원, 고위 임직원 및 관리자, 농업, 임업 및 어업 숙련 종사 제외이고 학력이 대졸이상 제외이고 거주지역이 어촌지역, 기타지역제외이고 연령이 76세 이상인 응답집단은 수돗물의 음용수 대책에 대하여 상수도 시설

의 대폭적 개선이 60%로 수돗물의 음용수 대책에 대하여 상수도 시설의 대폭적 개선에 더 비중을 주는 것으로 나타났다.

환경오염의 주체에 대한 의사결정나무 모형은 <그림 6>과 같다.

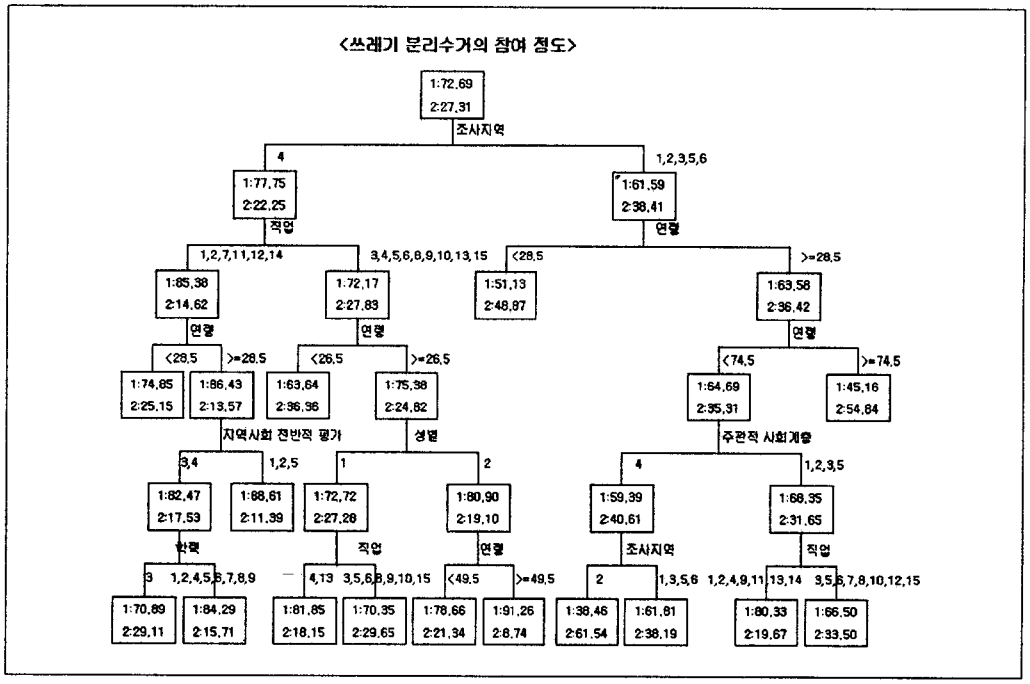


<그림 6> 환경오염의 주체에 대한 의사결정나무 모형

<그림 6>에서 살펴보면 최 상위 노드에서 최 상위 노드에서 환경오염의 주체에 대하여 1.모두 68.63% 2. 기업체, 일반 소비자 및 관광행락객, 농어민 이 31.37%로 전체적으로 68.63%의 응답자들이 모두라고 생각한다. 세부적으로 살펴 보면, 거주지역이 농촌지역, 어촌지역 제외이고 지역의 전반적 평가가 좋음과 보통이고 연령이 44세 이하인 응답집단은 환경오염의 해결주체로서 모두 75.36%로 환경오염의 주체로서 모두에 더 비중을 주는 것으로 나타났다. 그리고 거주지역이 농촌지역, 어촌지역 제외이고 지역의 전반적 평가가 좋음과 보통이고 연령이 45세 이상이고 성별이 여성이고 학력이 초졸이하, 고졸, 전문대졸, 대학원이상인 응답집단은 환경오염의 주체로서 모두 77.09%로 환경오염의 해결주체로서 모두에 더 비중을 주는 것으로 나타났다. 또한 거주지역이 농촌지역, 어촌지역 제외이고 지역의 전반적 평가가 좋음과 보통을 제외하고 직업이 의회의원, 고위임직원 및 관리자, 전문가, 기능원 및 관련 기능 종사자, 학생, 기타인 응답집단은 환경오염의 주체로서 모두 76.42%로 환경오염의 주체로서 모두에 더 비중을 주는 것으로 나타났다.

쓰레기 분리수거의 참여 정도에 대한 의사결정나무 모형은 <그림 7>과 같다. 최 상위 노드에서 쓰레기 분리수거 참여 정도에 대하여 1.높다 72.69%, 2. 낮다 27.31%로 전체적으로 72.69%의 응답자들이 쓰레기 분리수거 참여 정도에 대하여 전체적으로 높은 호응도를 보이고 있다. 세부적으로 살펴보면, 거주지역이 주거지역이고 직업이 의회

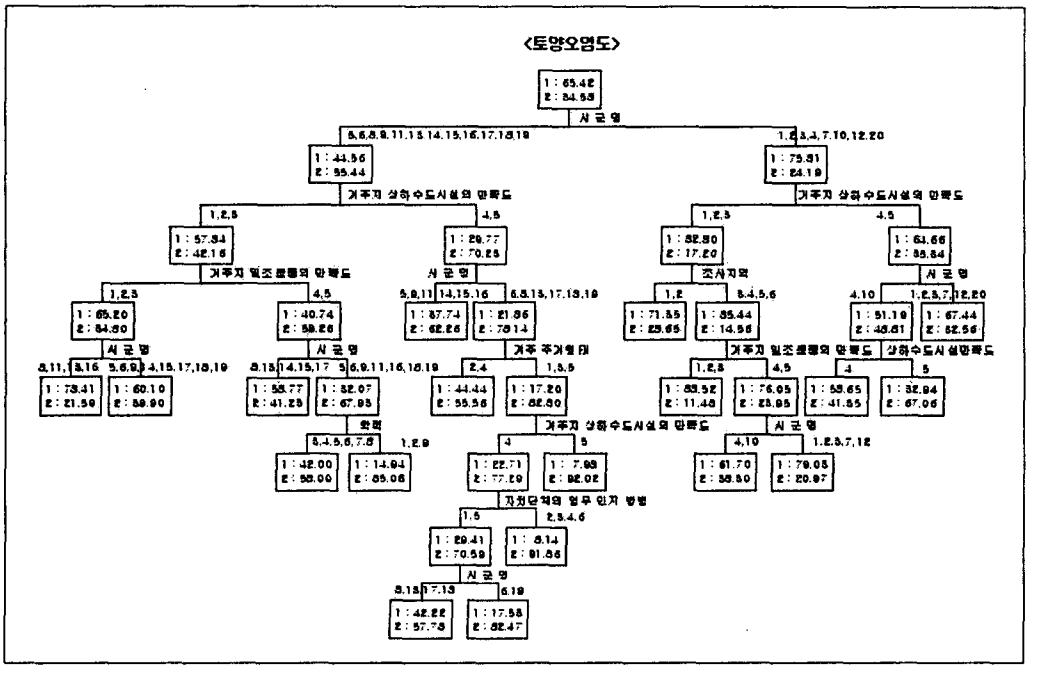
의원, 고위임직원 및 관리자, 전문가, 농업, 임업 및 어업 숙련 종사자, 군인, 가정주부, 학생이고 연령이 29세 이상이고 지역사회 전반적인 평가가 보통과 나쁨 제외인 응답집단은 쓰레기 분리수거 참여 정도에 대하여 높다 88.61%의 응답과 거주지역이 주거지역이고 직업이 의회의원, 고위임직원 및 관리자, 전문가, 농업, 임업 및 어업 숙련 종사자, 군인, 가정주부, 학생제외이고 연령이 50세 이상이고 성별이 여성인 응답집단은 쓰레기 분리수거 참여 정도에 대하여 높다 91.26%로 쓰레기 분리 수거의 참여 정도가 아주 높은 것으로 나타났다. 반면에 거주지역이 주거지역 제외 이고 연령이 74세 이하이고 주관적 사회계층이 중하류층 이고 거주지역이 어촌인 응답 집단은 쓰레기 분리수거 참여 정도에 대하여 낮다 61.54%로 쓰레기 분리수거의 참여 정도가 낮은 것으로 나타났다.



<그림 7> 쓰레기 분리수거의 참여 정도 의사결정나무 모형

녹색제품의 구입 정도에 대한 의사결정나무 모형은 <그림 8>과 같다. 최 상위 노드에서 녹색 제품의 구입 여부에 대하여 1.구매한다 40.44% 2. 구매하지 않는다 59.56%로 전체적으로 59.56%의 응답자들이 녹색제품을 구입 여부에 대하여 전체적으로 부정적인 시각을 보이고 있다. 세부적으로 살펴보면, 학력이 중졸이하이고 연령이 61세 이하이고 주관적 사회계층이 중하류층 이하이고 직업이 전문가, 판매종사자, 농업, 임업 및 어업 숙련 종사자, 장치, 기계조작 및 조립종사자, 단순노무종사자, 가정주부, 무직, 기타인 응답 집단은 녹색제품 구매하지 않는다 82.86%의 응답과 학력이 중졸이하고 연령이 62세 이상이고 주관적 사회계층이 상류층, 중하류층이하인 응답 집단은 녹색제품 구매하지 않는다 88.65%로 녹색제품 구매 여부에 대하여 부정적인 시각을 가지

고 있다. 반면에 학력이 전문대졸, 대졸, 대학원 이상이고 지역사회의 전반적인 평가가 좋음 이상이고 직업이 전문가, 기술공 및 준전문가, 사무종사자, 판매종사자, 단순노무종사자, 군인, 가정주부, 학생, 기타이고 연령이 35세 이상인 응답 집단은 녹색제품 구매한다 70.53%로 녹색제품 구매 여부에 대하여 긍정적인 시각을 가지고 있다.



<그림 8> 녹색제품의 구입 정도 의사결정나무 모형

5. 결론

본 논문에서는 환경 의식자료에 대하여 데이터마이닝의 기법 중 의사결정나무의 모형화 방안에 대하여 연구하고 2002년 조사된 경상남도 사회지표 조사의 자료에 대하여 의사결정나무 기법을 적용하여 모형을 구축하고 구축된 모형을 분석을 분석하였다. 환경 의식자료의 의사결정나무 모형화를 통하여 데이터 속에 잠재되어 있는 정보를 추출하여 환경 의식 자료에 대하여 도민들의 의식을 보다 심층적으로 세분화 할 수 있었으며, 쉽게 드러나지 않는 유용한 정보를 추출할 수 있었다. 향후 더욱 다양한 환경 관련 자료에 대하여 의사결정 나무 기법을 적용하여 데이터 속에 잠재되어 있는 다양하고 세분화된 정보를 추출할 수 있을 것이다.

참고문헌

1. 김정태, 정진도, 김광석(2003), 여름철 충청남도 서북부 지역에서의 대기오염물

질 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(2003), 2004년도 환경정보화촉진시행계획.
13. 환경부(2003a), 환경통계연감.
14. 환경부(2003b), 대기환경연보.
15. 환경부(2004), 환경백서.
16. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and regression trees*, Wadsworth, Belmont.
18. Hartigan, J.A. (1975), *Clustering Algorithms*, New York, John Wiley & Sons, Inc.
20. Loh, W.Y and Shin, Y,S(1997). *Split Selection Methods for Classification Tree*, Statistica Sinica. 7, 815-840.
21. Quinlan, J.R. (1993), *C4.5 Programs for Machine Learning*. San Mateo, Morgan Kaufmann.