Journal of the Korean Data & Information Science Society
2012, **23**(5), 983–992

http://dx.doi.org/10.7465/jkdi.2012.23.5.983
한국데이터정보과학회지

사회지표조사에서의 3단계 복합 데이터마이닝의 적용 방안

조광현[1] · 박희창[2]

[1]창원대학교 유아교육학과 · [2]창원대학교 통계학과

접수 2012년 8월 31일, 수정 2012년 9월 19일, 게재확정 2012년 9월 23일

요약

사회지표조사는 주민들이 생각하는 사회 상태를 총체적으로 파악할 수 있는 조사로서 다양한 시책 개발에 있어 지역의 여론을 반영할 수 있는 장점이 있다. 사회지표조사는 사회 변화를 알 수 있는 중요한 척도라고 할 수 있으며, 많은 지자체 (서울시, 인천시, 부산시, 울산시, 경상남도 등)에서 많은 예산과 시간을 들여 조사를 실시하고 있다. 그러나 조사에 대한 분석 결과가 기초통계 분석 위주로 되어 있어 실제 사회지표조사 자료를 제대로 활용하고 있지 못하고 있는 실정이므로 데이터마이닝 등의 다양한 방법의 적용이 필요하다. 이에 본 논문에서는 사회지표조사의 효율적인 분석을 위하여 새로운 데이터마이닝 방법론을 제시하고자 한다. 본 논문에서는 매개연관성규칙, k-평균 군집분석, 의사결정나무를 순차적으로 적용하는 3단계 복합 데이터마이닝의 적용 방법을 제안하며, 이를 2010년에 조사된 경상남도 사회지표조사 자료에 적용하고자 한다.

주요용어: 군집분석, 데이터마이닝, 매개연관성 규칙, 사회지표조사, 의사결정나무.

1. 서론

사회지표조사 (society indicator survey)는 주민들이 생각하는 사회 상태를 총체적이고 집약적으로 나타낼 수 있는 조사이다. 사회지표조사는 주민 생활의 양적 측면은 물론 질적 측면까지도 측정하기 때문에 사회의전반적 생활수준 및 의식을 파악할 수 있으므로 지역의 여론을 다양한 시책에 반영할 수 있는 이점이 있다. 지역여론을 반영한 시책은 지역개발에 대한 주민들의 의견과 정부의 방침을 상호 조화롭게 하여 지방자치제의 주된 목적을 효율적으로 달성할 수 있게 해준다.또한 지역여론을 기초한 시책은 전문성과 연계되어 한정된 투자재원을 효율적으로 배분할 수 있다. 이러한 시대적 요구와 지방자치제의 효과적 성숙을 위하여 서울시를 비롯한 인천시, 부산시, 울산시, 경상남도 등의 여러 지자체에서는 매년 사회지표조사를 통하여 주민들의 의식을 파악하고 있다. 이와 같이 사회지표조사는 변화하는 역사적 흐름 속에서 우리가 처해 있는 사회적 상태를 종합적으로 나타냄으로써 사회구성원들의 삶의 질을 전반적으로 파악하고 사회변화를 포착할 수 있는 척도라고 할 수 있다. 또한 사회지표조사는 주민들의 생활 수준, 사회의 종합적 상태, 사회변화의 예측, 사회개발정책의 성과 등을 측정하는 데 이용되고 있는 중요한 조사라고 할 수 있다. 이에 각 지차체에서는 많은 예산과 시간을 들여 사회지표조사를 실시하고 있다. 그러나 사회지표조사 결과 보고서를 살펴보면 대부분의 분석 결과가 기초통계분석 위주로 되어 있어 실제 사회지표 조사 자료를 제대로 활용하고 있지 못하고 있는 실정이다. 이에 데이터마이닝 (data mining) 등의 다양한 통계 분석 방법의 적용이 필요하다.

[1] (641-773) 경상남도 창원시 사림동 9번지, 창원대학교 유아교육학과, 통계학 시간 강사.
[2] 교신저자: (641-773) 경상남도 창원시 사림동 9번지, 창원대학교 통계학과, 교수.
E-mail: hcpark@changwon.ac.kr

사회지표조사에서의 데이터마이닝의 적용에 관한 연구로는 국내적으로 연구가 미비한 실정이다. 이에 본 논문에서는 사회지표조사 자료에 대하여 보다 심층적인 분석을 실시하기 위하여 새로운 데이터마이닝 방법론을 제시하고자 한다. 데이터마이닝은 방대한 양의 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 의미하며, 군집분석 (cluster analysis), 연결 분석 (link analysis), 판별 분석 (discrimination analysis), 연관성규칙 (association rule), 의사결정나무기법 (decision tree), 신경망모형 (neural network) 등의 다양한 분석 기법이 있다. 데이터마이닝의 여러 가지 기법 중 분류와 예측을 위하여 가장 많이 사용되는 방법이 의사결정나무기법이다. 일반적으로 의사결정나무 모형 생성 시, 관심대상이 되는 목표변수의 수가 많은 경우 여러 번의 모형 생성 과정을 거치게 된다. 이 때, 목표변수의 성향이 비슷할 경우 여러 개의 목표변수를 하나의 변수로 만들 수 있으면 의사결정나무 모형 생성 및 해석에 있어 효율적 일 수 있으며, 군집분석 방법을 통하여 여러 개의 변수를 통합하여 하나의 새로운 변수로 만들 수가 있다. 그러나 군집 분석 생성 시, 군집분석에 사용되는 변수의 수 및 변수의 관계에 따라서 군집의 결과가 다를 수가 있다. 이에 군집분석에 사용할 변수를 도출하기 위하여 연관성 규칙을 이용하고자 한다. 또한 연관성 규칙 시, 변수들 간의 간접적 관계가 존재 할 수 있으므로 매개변수 (intervening variable)를 도출할 수 있는 매개연관성규칙 (intervening association rule)을 적용하고자 한다. 이에 본 논문에서는 매개연관성규칙에 의하여 성향이 유사한 변수들을 도출하고 이 변수들을 이용하여 군집분석을 실시 한 후 의미 있는 군집분석 결과를 도출한다. 최종적으로 도출된 군집 분석 결과를 목표변수로 지정하여 의사결정나무 모형을 생성하고자 하므로 3단계 복합 데이터마이닝 방법론이라고 할 수 있다. 복합 데이터마이닝 방법에 대한 국내연구로는 Cho와 Park (2011a, 2011b, 2011c, 2012a, 2012b)이 있으며, 그리고 Park (2011a, 2011b) 등이 데이터마이닝과 관련된 연관규칙 평가 기준에 대한 연구를 진행하였다.

본 논문에서는 매개연관성규칙, 군집분석, 의사결정나무를 순차적으로 적용하는 방법을 제시한 후 2010년에 조사된 경상남도 사회지표조사 자료에 적용하고자 한다. 논문의 2절에서는 논문의 이론적 배경에 대하여 기술하고 3절에서는 연구방법에 대하여 기술하며 4절에서 실제자료 분석 결과를 제시한 후, 5절에서 결론을 맺고자 한다.

2. 이론적 배경

2.1. 매개연관성규칙

연관성규칙은 항목 집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 반영하는 규칙으로서 Agrawal 등 (1993)에 의해 처음 소개되어 졌다. 일반적으로 독립변수와 종속변수 간의 연관성규칙 생성 시, 우연히 매개변수와 연결됨으로써 관련성이 있는 것으로 나타나는 경우가 발생할 수 있다. 독립변수와 종속변수 사이에 매개변수가 존재하는 경우 두 변수 간에는 실제적인 관련성이 없으나 매개변수에 의하여 관련성이 있는 것으로 나타날 수 있다. 이 경우 두 변수간의 관련성을 분석한다면 잘못된 해석을 내릴 수 있다. 이때 매개변수란 독립 변수와 종속 변수 사이에서 독립 변수의 결과인 동시에 종속 변수의 원인이 되는 변수를 의미하며, 매개변수가 존재하는 경우 독립변수와 종속변수간의 직접적인 관련성은 없는 것으로 판단한다. Cho와 Park (2011c)은 연관성 규칙을 이용하여 매개변수를 추출하는 방법에 대하여 연구한 바 있고, 이 방법을 매개연관성규칙이라고 명하였으며, 그 조건은 다음과 같다. 다음의 4가지 조건이 만족하면, 매개변수에 의한 전항변수와 후항변수간의 규칙은 큰 의미가 없는 것으로 판단한다.

[조건 1] Y (후항변수)와 X_1 (전항변수)에 대한 연관성규칙의 결과가 지정된 최소 지지도와 최소 신뢰도보다 커야 한다.

[조건 2] X_1과 X_2 (매개변수)에 대한 연관성규칙의 결과가 지정된 최소 지지도와 최소 신뢰도보다 커야 한다.

[조건 3] X_1 및 X_2와 Y와의 연관성규칙의 결과가 지정된 최소 지지도와 최소 신뢰도보다 커야 한다.

[조건 4] X_1 및 X_2와 Y와의 연관성규칙의 신뢰도가 Y와 X_1에 대한 연관성규칙의 신뢰도보다 커야 한다.

2.2. 군집분석

군집분석은 다양한 특성을 지닌 관찰대상을 유사성을 바탕으로 동질적인 집단으로 분류하는데 쓰이는 기법이다. 즉, 데이터의 물리적 혹은 추상적 객체를 비슷한 객체군으로 묶는 과정이라 할 수 있다. 군집분석의 기본 목적은 관찰대상이 되는 개체들의 집합을 여러 개의 자연스러운 군집으로 분류하는 데 있다. 분류된 군집들은 상호 배타적이어서 한 군집에 속한 개체들은 서로 유사한 성질을 갖지만, 이들은 다른 군집에 속한 개체들과는 서로 다른 성질을 가지고 있다. 유사성의 측정은 개체의 특성에 대한 측정치들을 거리로 환산하여 측정하게 되며, 유클리디안 거리 (Euclidean distance), 유클리디안 제곱거리 (squared Euclidean distance), 마할라노비스 (Mahalanobis distance), 민코우스키 거리 (Minkowski distance) 등 네 가지 방식이 있다. 일반적으로 이중에서 유클리디안 제곱거리를 가장 많이 사용한다. 유클리디안 거리는 변수값들의 차이를 제곱하여 합산한 거리, 다차원공간에서 직선최단 거리를 말한다.

군집분석의 방법에는 분할 군집법, 계층적 군집법, 밀도에 의한 군집법, 그리드에 의한 군집법, 모형에 의한 군집법 등이 있다. 이들 중에서 분할 군집법은 데이터들을 임의의 부분집합으로 분할을 한 후 데이터들을 유사한 그룹으로 재배치하여 분할하는 것을 개선하려고 하는 군집방법이다. 분할 군집법 중 k-평균 군집 분석이 가장 많이 사용된다. k-평균 군집분석은 MacQueen (1967)에 의해 처음 소개되어진 분할군집법의 일종으로 데이터들을 k개의 군집으로 임의로 분할을 하여 군집의 평균을 대표값으로 분할해 나가는 방법으로 데이터들을 유사성을 바탕으로 재배치를 하는 방법이다.

2.3. 의사결정나무

의사결정나무는 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. 그 동안의 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 의사결정나무 알고리즘에는 Hartigan (1975)에 의하여 제시된 CHAID (Chi-squared Automatic Interaction detecion), Breiman 등 (1984)에 의하여 제시된 CART (Classification and Regression Trees), Quinlan (1993)의 ID3을 기반으로 한 C5.0 등의 알고리즘 있으며, CHAID는 의사결정나무의 가장 오래된 알고리즘으로 분리기준으로 카이제곱통계량을 사용하고, CART는 분리기준으로 지니 지수를 사용하여 이지 분리를 수행하는 알고리즘이며, C5.0은 분리기준으로 엔트로피를 사용하여 다지 분리를 수행하는 알고리즘이다.

3. 연구 방법

본 논문은 효율적인 의사결정나무 생성을 위하여 매개연관성규칙, 군집분석, 의사결정나무를 순차적으로 적용하는 3단계 데이터마이닝 적용 방안이라고 할 수 있다. 본 논문에서 제안하는 연구

방법을 자세하게 설명하면 Figure 3.1과 같다.

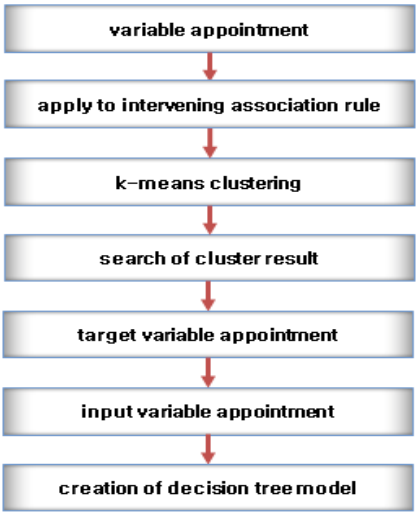

Figure 3.1 Application method

Figure 3.1을 자세하게 설명하면 다음과 같다.

[단계 1] 변수 지정
의사결정나무 모형을 생성하기 위하여 관심대상이 되는 변수들을 지정한다.

[단계 2] k-평균 군집분석에 사용할 변수 추출을 위한 매개연관성규칙 적용
변수들의 관련성을 알아보기 위하여 매개연관성규칙을 적용한다. 매개연관성규칙에 의하여 도출된 변수들의 관련성은 k-평균 군집분석의 변수 축소에 사용된다. 세부 과정은 다음과 같다.

1) 연관성 규칙 생성 : 지정된 변수들에 대하여 최소지지도, 최소신뢰도, 향상도를 지정하여 연관성 규칙을 생성한다.

2) 매개변수 지정 : 생성된 규칙에 대한 매개변수가 존재하는 가를 파악하기 위하여 매개변수를 지정한다.

3) 매개변수 조건 성립 파악 : 지정된 매개변수에 대하여 실제로 매개변수 조건이 성립하는 가에 대하여 파악한다. 매개변수 조건은 2.1절에서 설명한 네 가지 조건을 만족해야 하며 네 가지 조건을 모두 만족하는 경우 지정된 전항변수는 의미가 없다고 판단한다. 조건 성립을 위하여 다음의 단계를 거친다.

   3-1) 전항변수와 매개변수와의 연관성 규칙 생성

   3-2) 전항변수 및 매개변수와 후항변수와의 연관성 규칙 생성

   3-3) 기존의 연관성 규칙 결과와 3-2단계의 연관성 규칙 결과 비교

[단계 3] 변수 축소를 위한 k-평균 군집분석 실시
앞의 매개연관성규칙에서 추출된 연관성규칙을 바탕으로 실제적으로 의사결정나무 분석에서 목표변수로 사용된 변수를 추출하기 위하여 k-평균 군집분석을 실시한다. 세부 과정은 다음과 같다.

1) 변수 선정 : k-평균 군집분석에 사용할 변수를 선정한다. 변수는 앞의 매개연관성규칙에서 추출된 관련성 있는 변수들을 지정한다.

2) k-평균 군집분석 실시 : 군집의 특성이 명확하게 파악되는 군집을 도출하기 위하여 군집의 수를 2개∼5개로 하여 k-평균 군집분석을 실행한다. k-평균 군집분석은 군집의 수를 연구자가 임의로 지정해야 한다는 단점이 있어 본 논문에서는 군집의 수를 다양하게 지정하여 군집분석을 실시한다.

[단계 4] 군집 비교 및 결정

앞에서 실시한 4개의 군집분석의 결과를 비교한다. 군집분석을 비교한 후 군집의 특성을 가장 명확하게 해주는 군집의 수를 결정한다. 또한 최종 결정된 군집 결과를 의사결정나무 분석의 목표변수로 사용하기 위하여 군집 결과를 하나의 변수로 저장한다.

[단계 5] 목표변수 지정

의사결정나무 모형을 생성하기 위하여 목표변수를 지정한다. 목표변수는 앞서 변수로 저장한 k-평균 군집분석의 결과를 지정한다.

[단계 6] 입력변수 지정

의사결정나무 모형을 생성하기 위하여 입력변수를 지정한다. 입력변수는 관심대상이 되는 목표변수에 대한 응답자들의 분류를 알아보기 위함이므로 일반적으로 인구통계학적 속성을 지정한다.

[단계 7] 의사결정나무 모형 생성

지정된 목표변수와 입력변수와의 의사결정나무 모형을 생성한다. 모형생성 시, 목표변수와 입력변수의 형태에 따라 CHAID, CART, C5.0의 알고리즘 중 하나를 선택한다. 일반적으로 생성된 의사결정나무 모형이 복잡해지면 해석이 어렵게 된다. 본 논문에서는 비교적 모형이 간단한 CART 알고리즘을 사용하여 의사결정나무 모형을 생성하며, 의사결정나무의 모형화 보다는 응답자의 분류에 주안점을 두기 위함이므로 훈련자료와 평가자료로 나누지 않고 모형을 생성한다.

4. 자료 분석

본 절에서는 논문에서 제안하는 방법을 2010년에 조사된 경상남도사회지표조사 자료에 적용하기로 한다. 경상남도는 1992년부터 매년 설문 조사를 통하여 도민의식을 파악하고 있으며, 당해 년도에 필요하고 부적절한 몇 몇 첨삭항목을 제외하고는 3년 주기로 재조사를 하여 의식변화를 비교 분석해오고 있다. 2010년 경남사회지표조사에서는 조사대상 모집단을 경남 18개 시·군에 거주하는 17세 이상의 모든 도민과 가구로 하였다. 경남사회지표조사의 자료구조는 크게 일반사항 (인구통계학적 문항)과 도민의식조사부문으로 나누어져 있다. 일반사항은 조사응답자의 연령, 성별, 학력, 가구주와의 관계, 결혼 유무, 직업 등으로 구성되어 있으며, 도민의식조사부문은 소득·소비, 고용·노사, 교육, 보건·체육, 주택, 환경, 사회, 정보화, 문화·여가, 그리고 안전부문 등으로 구성되어 있다. 이중 환경부문에 대하여 Table 4.1과 같이 8개 항목에 대한 오염을 조사하고 있다.

Table 4.1 Variables of pollution

no	variable	form	no	variable	form
1	waterworks pollution	5 likert scale	5	atmosphere pollution	5 likert scale
2	sewerage pollution	5 likert scale	6	soil pollution	5 likert scale
3	noise pollution	5 likert scale	7	river pollution	5 likert scale
4	stink pollution	5 likert scale	8	sea pollution	5 likert scale

Table 4.1에서 보는 바와 같이 오염에 관하여 상수도, 하수도, 소음, 악취, 대기, 토양, 하천, 해양 오염 등 총 8개의 문항이 있다. 각 오염에 대하여 응답자들의 속성을 분류하기 위하여 의사결정나무 생성하면 총 8개의 모형을 생성해야 한다. 그러나 8개의 오염에 대한 문항을 자세하게 살펴보면 속성이 유사하거나 비슷한 경향이 나타나는 문항이 존재할 수 있다. 만일 비슷한 속성을 가지는 문항들을 하나의 변수로 축소할 수 있다면 의사결정나무 모형의 생성 및 해석에 있어 효율적일 수 있다. 이에 본 절에서는 오염의 8개 문항에 대하여 본 논문에서 제안하는 3단계 복합 데이터마이닝 방법을 적용하고자 한다. 우선 관심대상이 되는 변수는 앞서 설명한 오염에 대한 8개의 변수이고 각 변수들의 관련성을 알아보기 위하여 매개연관성규칙을 적용하였다. 우선 8개 변수들에 대한 연관성 규칙 (최소 지지도 : 10, 최소 신뢰도 : 70, 향상도 : 1) 결과는 Table 4.2 및 Table 4.3과 같다.

Table 4.2 Result of association rule (1)

no	antecedent	consequent	support	confidence
1	waterworks	sewerage	35.70	94.93
2	river	sea	27.01	86.26
3	waterworks	river	35.39	80.13
4	waterworks	sea	27.01	79.26
5	sewerage	river	35.39	77.12
6	sewerage	sea	27.01	74.97

Table 4.3 Result of association rule (2)

no	antecedent	consequent	support	confidence
1	atmosphere	soil	41.53	90.65
2	atmosphere	noise	31.60	86.01
3	atmosphere	stink	36.19	85.82
4	soil	stink	36.19	83.55
5	soil	noise	31.60	82.40
6	stink	noise	31.60	80.72

Table 4.2 및 Table 4.3의 연관성 규칙 결과를 살펴보면 다음과 같다. 우선 Table 4.2의 경우 상수도 오염, 하수도 오염, 하천 오염, 해양 오염의 4가지 항목에 대하여 관련성이 높게 나타나는 것을 알 수 있으며, 이 문항들은 수질 오염이라고 할 수 있다. 또한 Table 4.3의 경우 소음 오염, 악취 오염, 대기 오염, 토양 오염의 4가지 항목에 대하여 관련성이 높게 나타나는 것을 알 수 있으며, 이 문항들은 토양 · 대기 오염이라고 할 수 있다.

연관성 규칙에 의하여 나타나는 규칙들 중 실제로 매개변수가 존재하는 가를 파악하기 위하여 매개연관성규칙을 적용하였다. 적용 결과 Table 4.4와 같이 하수도 오염과 해양 오염 사이에 하천 오염이 매개 변수 역할을 하는 것으로 나타났다. 이에 매개 변수인 하천오염에 의하여 하수도 오염은 의미없는 변수로 판단하여 k-평균 군집분석에서는 사용하지 않는다.

Table 4.4 Result of intervening association rule

no	antecedent	intervening	consequent	confidence
1	sewerage	-	sea	79.26
2	sewerage	river	-	80.13
3	sewerage	river	sea	86.71

이에 매개연관성규칙의 결과를 바탕으로 변수 축소를 위하여 k-평균 군집분석을 실시한다. k-평균 군집분석에서는 수질오염의 3문항 (상수도 오염, 하천 오염, 해양 오염)과 토양 · 대기 오염의 4문항 (소음 오염, 악취 오염, 대기 오염, 토양 오염)에 대하여 군집분석을 실시하였다. k-평균 군집분석 시,

군집의 수를 2개에서 5개로 다양하게 지정하여 분석을 실시한 결과 Table 4.5 및 Table 4.6과 같이 군집의 수를 2개로 지정하였을 때, 군집의 특성이 명확하게 구분되었다.

Table 4.5 Result of k-means clustering (water pollution)

cluster	cluster 1	cluster 2
waterworks	3.53	2.73
river	3.48	2.17
sea	3.50	2.37
case	6,252	1,075

Table 4.6 Result of k-means clustering (earth and air pollution)

cluster	cluster 1	cluster 2
noise	3.80	2.58
stink	3.93	2.78
atmosphere	4.11	2.89
soil	4.06	2.98
case	4,036	5,959

Table 4.5를 살펴보면, 군집 1의 응답자가 군집 2의 응답자 보다 상수도 오염, 하천 오염, 해양 오염의 수치가 높은 것 (수치가 높을수록 만족함이 높음)을 알 수 있어 군집 1의 집단은 오염에 긍정적인 집단으로 분류할 수 있고 군집 2의 집단은 오염에 부정적인 집단으로 분류할 수 있다. 또한 Table 4.6을 살펴보면, 군집 1의 응답자가 군집 2의 응답자 보다 소음 오염, 악취 오염, 대기 오염, 토양 오염의 수치가 높은 것을 알 수 있어 군집 1의 집단은 오염에 긍정적인 집단으로 분류할 수 있고 군집 2의 집단은 오염에 부정적인 집단으로 분류할 수 있다.

다음으로 변수 축소로 나타난 수질 오염 결과와 토양 · 대기 오염 결과를 목표변수로 지정하여 의사결정나무 모형을 생성한다. 의사결정나무 모형 생성 시 입력변수로는 나이, 성별, 학력, 직업, 건강상태, 월평균 소득의 6개 인구통계학적 문항을 사용하였다. 생성된 의사결정나무 모형은 Figure 4.1 및 Figure 4.2와 같다. Figure 4.1의 수질 오염에 대한 의사결정나무 모형을 살펴보면 다음과 같다. 수질 오염에 대하여 나이가 30대 이하이고 건강상태가 양호하며, 수입이 200만원 이하인 주민들은 수질 오염에 대하여 전체 보다 긍정적인 응답을 하고 있으나, 나이가 30대 이하이고 건강상태가 좋지 않은 주민들과 나이가 30대 이하이고 건강상태가 양호하나 수입이 200만원 이상인 주민들은 수질 오염에 대하여 전체 보다 부정적인 응답을 하고 있어 차이를 보이고 있다. 또한 나이가 40대 이상이고 건강상태가 양호하며, 직업이 관리자 및 전문직인 주민들은 수질 오염에 대하여 전체 보다 부정적인 응답을 하고 있으나, 나이가 40대 이상이고 건강상태가 양호하나 직업이 생산직인 주민들은 수질 오염에 대하여 전체 보다 긍정적인 응답을 하고 있어 차이를 보이고 있다.

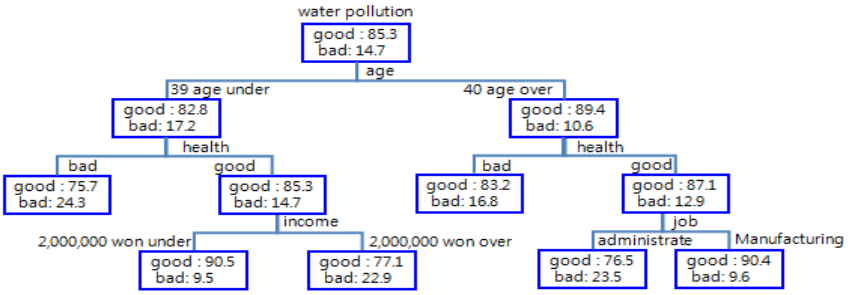

Figure 4.1 Result of decision tree model (water pollution)

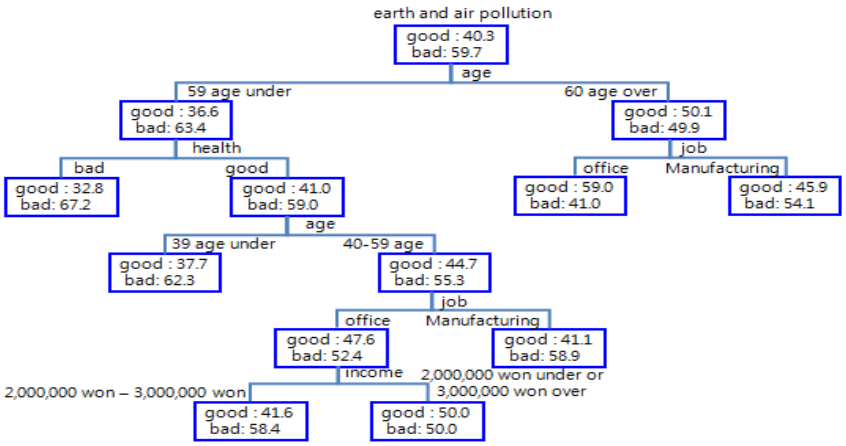

Figure 4.2 Result of decision tree model (earth and air pollution)

Figure 4.2의 토양 · 대기 오염에 대한 의사결정나무 모형을 살펴보면 다음과 같다. 토양 · 대기 오염에 대하여 나이가 40대와 50대이고 건강상태가 양호하고 직업이 사무직이며, 수입이 200만원~300만원인 주민들은 토양 · 대기 오염에 대하여 전체 보다 긍정적인 응답을 하고 있으나, 나이가 50대 이하이고 건강상태가 좋지 않은 주민들은 토양 · 대기 오염에 대하여 전체 보다 부정적인 응답을 하고 있어 차이를 보이고 있다. 또한 나이가 60대 이상이고 직업이 사무직인 주민들은 토양 · 대기 오염에 대하여 전체 보다 긍정적인 응답을 하고 있는 것으로 나타났다.

5. 결론

현재 사회의전반적 생활수준 및 의식을 파악하여 지역의 여론을 다양한 시책에 반영하기 위하여 각 지자체에서는 사회지표조사를 실시하고 있다. 실제로 사회지표조사는 사회 변화를 알 수 있는 중요한 척도라고 할 수 있으며, 우리가 처해 있는 사회적 상태를 종합적으로 나타냄으로써 사회구성원들의 삶의 질을 전반적으로 파악할 수 있다. 그러나 각 지자체에서 많은 예산과 시간을 들여 사회지표조사를 실시하고 있으나, 조사 자료의 분석이 단순 통계분석에 그쳐 실제 사회지표 조사 자료를 제대로 활용하고 있지 못하고 있는 실정이다. 이에 본 논문에서는 효율적인 의사결정나무 생성을 위하여 매개연관성규칙, 군집분석, 의사결정나무를 순차적으로 적용하는 3단계 복합 데이터마이닝 적용 방법을 제시하였다. 3단계 복합 데이터마이닝은 매개연관성규칙에 의한 변수들 간의 관계를 파악한 뒤, 이를 바탕으로 k-평균 군집분석을 통하여 여러 개의 변수들을 축약하고 이 축약된 결과를 이용하여 의사결정나무 분석을 실시하는 방법을 제안하였다.

실제 2010년 조사된 경상남도 사회지표조사 자료에 대하여 본 논문에서 제안하는 방법을 적용한 결과, 관심대상이 되는 8개의 오염 변수를 수질 오염, 토양 · 대기 오염의 2개의 변수로 축약할 수 있었다. 즉, 원래 관심대상이 되는 8개 변수 각각에 대한 의사결정나무 모형을 생성해야 하나, 본 논문에서 제안하는 방법을 이용하면 2개의 의사결정나무 모형만으로도 해석이 가능하므로 의사결정나무 모형 생성 및 해석에 있어 효율적이라고 할 수 있다. 추후 연구 과제로 변수들 간의 관계 및 변수 축약에 있어 꼭 연관성 규칙 및 군집분석을 사용해야 하는 것은 아니므로 변수들 간의 관계 파악 및 다양한 변수들을 축약하여 새로운 변수로 추출하는 방법에 대하여 여러 가지 다양한 분석 방법을 접목해 볼 필요성이 있다.

참고문헌

Agrawal, R., Imielinski, R. and Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD Conference on Management of Data*, 207-216.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and regression trees*, Wadsworth and books, California.

Cho, K. H. and Park, H. C. (2011a). A study on decision tree creation using intervening variable. *Journal of the Korean Data & Information Science Society*, **22**, 671-678.

Cho, K. H. and Park, H. C. (2011b). A study on removal of unnecessary input variables using multiple external association rule. *Journal of the Korean Data & Information Science Society*, **22**, 877-884.

Cho, K. H. and Park, H. C. (2011c). A study on insignificant rules discovery in association rule mining. *Journal of the Korean Data Analysis Society*, **22**, 81-88.

Cho, K. H. and Park, H. C. (2012a). A study on association rule creation by marginally conditional variables. *Journal of the Korean Data & Information Science Society*, **23**, 121-129.

Cho, K. H. and Park, H. C. (2012b). A study on decision tree creation using marginally conditional variables. *Journal of the Korean Data & Information Science Society*, **23**, 299-307.

Hartigan, J. A. (1975). *Clustering algorithms*, John Wiley & Sons, New York.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281-297.

Park, H. C. (2011a). Proposition of negatively pure association rule threshold. *Journal of the Korean Data & Information Science Society*, **22**, 179-188.

Park, H. C. (2011b). The proposition of attributably pure confidence in association rule mining. *Journal of the Korean Data & Information Science Society*, **22**, 235-243.

Quinlan, J. R. (1993). *C4.5 programs for machine learning*, Morgan Kaufmann Publishers, San Francisco.

Journal of the Korean Data & Information Science Society
2012, **23**(5), 983–992

http://dx.doi.org/10.7465/jkdi.2012.23.5.983
한국데이터정보과학회지

A study on 3-step complex data mining in society indicator survey

Kwang-Hyun Cho[1] · Hee-Chang Park[2]

[1]Department of Early Childhood Education, Changwon National University
[2]Department of Statistics, Changwon National University

Received 31 August 2012, revised 19 September 2012, accepted 23 September 2012

Abstract

Social indicator survey can identify the state of society as a whole. When we create a policy, social indicator survey can reflect the public opinion of the region. Social indicator survey is an important measure of social change. Social indicator survey has been conducted in many municipalities (Seoul, Incheon, Busan, Ulsan, Gyeongsangnam-do, etc.). But, the result of social indicator survey analysis is mainly the basic statistical analysis. In this study, we propose a new data mining methodology for effective analysis. We propose a 3-step complex data mining in society indicator survey. 3-step complex data mining uses three data mining method (intervening association rule, clustering, decision tree).

Keywords: Clustering, data mining, decision tree, intervening association rule, society indicator survey.

[1] A part-time lecturer, Department of Early Childhood Education, Changwon National University, Changwon, Gyeongnam 641-773, Korea.
[2] Corresponding author: Professor, Department of Statistics, Changwon National University, Changwon, Gyeongnam 641-773, Korea. E-mail: hcpark@changwon.ac.kr