사회복지 지표조사 자료의 데이터 마이닝 기법 적용 방안 모델

김창기*, 윤의정**, 이상도**, 서정민°
한국교통대학교 사회복지학과*, 충북사회복지협의회**, 충북융합복지연구소°
e-mail: cgkim@ut.ac.kr*, cpcsww@chol.com**, nextto2@hanmail.net**, jmseo@kku.ac.kr°

A Model on Data Mining Technique in Society Indicator Survey on Social Service

Chang Gi Kim*, Yun Eui Jeong**, Lee Sang Do**, Jeong Min Seo°
*Dept. of Social Welfare, Korea Nat'l Univ. of Transportation
**Chungbuk Council on Social Welfare
°Convergence Social Welfare Institute

● 요 약 ●

사회복지 지표조사는 주민들의 사회복지에 관한 다양한 상태를 전체적으로 파악할 수 있는 조사로 다양한 복지정책 개발에 있어 지역 주민들의 욕구를 반영할 수 있는 장점이 있다. 따라서 사회복지 지표조사는 복지 욕구를 알 수 있는 중요한 척도의 기준이라 할 수 있어 많은 지자체 및 정부기관에서 예산과 시간을 들여 조사를 실시하고 있다. 그러나 조사에 대한 분석 결과가 기초적인 통계 분석 위주로 되어 있어 실제 자료를 제대로 활용하지 못하고 있는 것이 현실이다. 이에 본 논문에서는 데이터 마이닝 기법을 이용한 분석 방법론을 제시한다. 또한 이를 충청북도의 중장기 사회복지 정책을 위한 지표조사에 적용하였다.

키워드: 사회복지(Social Service), 지표조사(Indicator Survey), 데이터 마이닝(Data Mining)

I. 서 론

사회복지 지표조사는 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. 이와 같이 사회복지 지표조사는 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 사회복지 지표조사는 주민들의 생활 수준, 사회의 종합적 상태, 사회변화의 예측, 사회개발정책의 성과 등을 측정하는데 이용되고 있는 중요한 조사라고 할 수 있다. 사회지표조사에서의 데이터마이닝의 적용에 관한 연구로는 국내적으로 연구가 미비한 실정이다. 이에 본 논문에서는 사회복지 지표조사 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 모형생성 시, 관심대상이 되는 목표변수의 수가 많은 경우 여러 번의 모형 생성 과정을 거치게 된다. 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 생성한다. 데이터 마이닝 방법에 대한 국내 연구로는 [1, 2, 3]이 있으며, 그리고 [4] 등이 데이터 마이닝과 관련된 연관규칙 평가 기준에 대한 연구를 진행하였다.

II. 관련연구

연관성 규칙은 항목 집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 반영하는 규칙으로서 [5]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. [1, 4]는 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 X1 (전항변수)에 대한 연관성규칙의 결과가 지정된 최소 지지도와 최소 신뢰도보다 커야 한다.

[조건 2] X1과 X2 (매개변수)에 대한 연관성규칙의 결과가 지정된 최소 지지도와 최소 신뢰도보다 커야 한다.

[조건 3] X1 및 X2와 Y와의 연관성규칙의 결과가 지정된 최소 지지도와 최소 신뢰도보다 커야 한다.

[조건 4] X1 및 X2와 Y와의 연관성규칙의 신뢰도가 Y와 X1에 대한 연관성규칙의 신뢰도보다 커야 한다.

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. 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 [6]에 의해 처음 소개되어진 분할군집법의 일종으로 데이터들을 k개의 군집으로 임의로 분할을 하여 군집의 평균을 대표값으로 분할해 나가는 방법으로 데이터들을 유사성을 바탕으로 재배치를 하는 방법이다.

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.

III. 마이닝 기법의 적용

본 논문은 효율적인 의사결정나무 생성을 위하여 매개연관성규칙, 군집분석, 의사결정나무를 순차적으로 적용하였다.

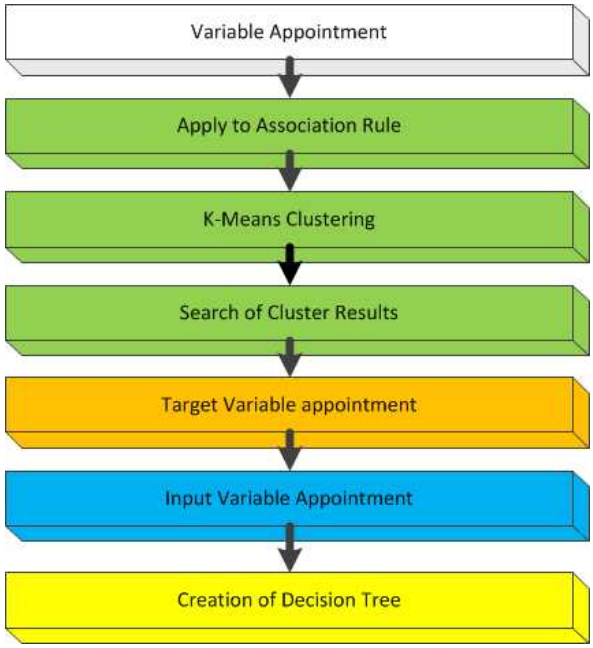

그림 298. 데이터 마이닝 적용 프로세서
Fig. 1. Application Processes

본 논문에서 제안하는 방법을 2014년 제3기 충청북도 중장기 지역사회복지계획을 위한 설문 및 욕구조사에 적용하였다. 본 조사는 2015년부터 2018년도까지의 충청북도 지역사회복지 계획을 위한 조사로 충북의 11개 자자체를 대상으로 충북에서 하고 있는 복지사업에 대한 인식 및 만족도, 충북에서의 특화된 복지사업 발굴, 현황을 조사하였다. 자료조사는 크게 일반사항(인구통계학적 문항)과 도민의 사회복지 의식조사의 두 부문으로 구성되어 있다. 사회복지 의식조사 부문은 이용경험, 이용 만족도, 정보획득 방법, 정보제공 방법 등의 21개 문항으로 구성되어 있다.

이중 6개의 조건(9988행복 나누미, 여성농업인 복지, 영유아 보육, 노인 일자리, 다문화가족, 노인장기/용양보험/돌봄서비스)로 구성된 5개의 주요 문항(이용경험, 이용 만족도, 분양별 중요도, 이용경험, 힘써야 할 복지 분야)을 이용하여 응답자들의 속성을 분류하기 위하여 의사결정나무를 생성하기 위해 비슷한 속성을 가지는 문항들을 하나의 변수로 축소하여 모형의 효율적인 생성 및 해석이 가능하도록 하였다.

우선 5개의 문항과 6개의 조건에 관해 변수들에 대한 연관성 규칙(최소 지지도: 10, 최소 신뢰도:70, 향상도:1)의 결과는 표 1과 같다.

표 1. 연관성 규칙 결과
Table. 1. Result of Association Rule

No	Ante.	Cons.	Supp.	Conf.
1	9988	9988	38.92	95.63
2	여성농업	영유아	37.65	93.26
3	영유아	영유아	42.23	95.35
4	노인일자리	9988	27.01	79.87
5	다문화	영유아	35.39	77.32
6	노인장기	9988	27.03	74.89

표 1의 결과를 살펴 보면 9988, 여성농업인, 영유아, 노인일자리, 다문화, 노인장기 등의 6가지 항목에 대하여 각각의 연관성은 9988, 여성농업인, 영유아로 축소할 수 있다. 연관성 규칙에 의하여 나타나는 규칙들 중 실제로 매개변수가 존재하는 가를 파악하기 위하여 표 1의 결과를 이용하여 매개 연관성 규칙을 적용하였다. 적용 결과 표 2, 표 3과 같이 여성농업인, 영유아, 다문화 사이에 영유아가 매개 변수의 역할을 하며, 9988 및 노인일자리, 노인장기의 사이에는 노인이 매개 변수 역할을 하는 것으로 나타났다.

표 2. 매개 연관성 규칙의 결과(1)
Table. 2. Result of Intervening Association Rule(1)

No	Ante.	Inte.	Cons.	Conf.
1	9988	-	9988	78.63
2	9988	노인일자리	9988	81.36
3	9988	노인장기	9988	87.58

표 2와 표 3에 9988과 영유아 사업은 의미 없는 매개 변수로 판단하여 K-평균 군집 분석에서는 사용하지 않는다.

표 3. 매개 연관성 규칙의 결과(2)
Table. 3. Result of Intervening Association Rule(2)

No	Ante.	Inte.	Cons.	Conf.
1	영유아	-	영유아	76.38
2	영유아	여성농업	영유아	82.15
3	영유아	다문화	영유아	84.98

표 2와 표 3의 결과를 바탕으로 변수 축소를 위하여 k-평균 군집분석을 실시한다. K-평균 군집분석 시, 군집의 수를 2개에서 5개로 다양하게 지정하여 분석을 실시한 결과 표 4와 같이 군집의 수를 2개로 지정하였을 때, 군집의 특성이 명확하게 구분되었다.

표 4. k-평균 군집분석의 결과
Table. 4. Result of k-means clustering

Ante.	Cluster 1	Cluster 2
9988	3.47	2.58
영유아	3.56	2.17
Case	1,036	213

표 4의 결과를 살펴보면 군집 1의 응답자가 군집 2의 응답자보다 9988과 영유아 복지사업에 관한 수치가 높은 것을 알 수 있다. 수치가 높으면 만족도가 높다. 즉, 군집 1은 긍정 집단으로 군집 2는 부정적 집단으로 분류가 가능하다. 이러한 결과를 이용하여 의사결정나무 모형을 생성한다. 의사결정나무 모형 생성 시 입력변수로는 나이, 성별, 학력, 직업, 건강상태, 월평균 소득의 6개 인구통계학적 문항을 사용하였다. 생성된 의사결정나무 모형은 그림 2, 그림 3과 같다.

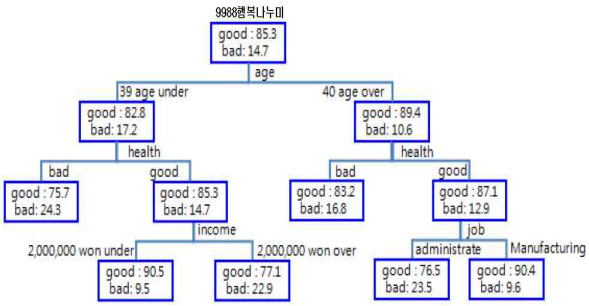

그림 299. 의사결정나무 모델 결과(9988행복나누미)
Fig. 2. Result of decision tree model(9988 Biz)

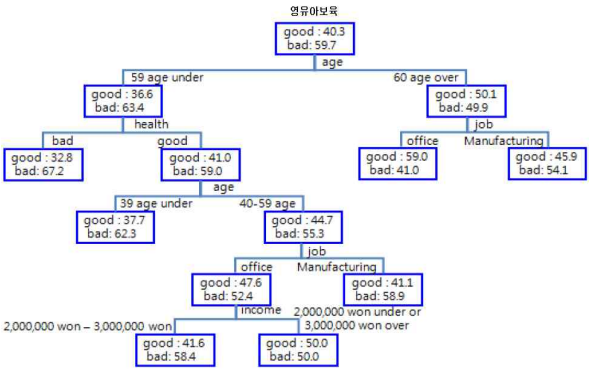

그림 300. 의사결정나무 모델 결과(영유아보육)
Fig. 3. Result of decision tree model(Infants Biz)

Ⅳ. 결 론

현재 사회의 전반적 복지수준 및 의식, 욕구를 파악하여 지역의 여론을 다양한 시책에 반영하기 위하여 각 지자체에서는 사회복지 지표조사를 실시하고 있다. 실제로 사회복지 지표 조사는 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 사회복지 지표 조사를 실시하고 있으나, 조사 자료의 분석이 단순 통계분석에 그쳐 실제 사회복지 지표조사 자료를 제대로 활용하고 있지 못하고 있는 실정이다. 이에 본 논문에서는 효율적인 의사결정나무 생성을 위하여 매개연관성규칙, 군집분석, 의사결정나무를 적용하는 데이터마이닝 적용 방법을 제시하였다. 제안한 방법은 데이터마이닝은 매개연관성규칙에 의한 변수들 간의 관계를 파악한 뒤, 이를 바탕으로 K-평균 군집분석을 통하여 여러 개의 변수들을 축소하고 이 축소된 결과를 이용하여 의사결정나무 분석을 실시하는 방법을 제안하였다. 실제 2014년 조사된 충청북도 중장기 복지계획 수립을 위한 사회복지 지표조사 자료에 대하여 본 논문에서 제안하는 방법을 적용한 결과, 관심대상이 되는 5개의 문항과 6개의 사회복지 변수를 9988행복나누미 사업과 영유아 보육의 2개의 변수로 축소할 수 있었다. 즉, 원래 관심대상이 되는 문항별 항목별 약 30개 변수 각각에 대한 의사결정나무 모형을 생성해야 하나, 본 논문에서 제안하는 방법을 이용하면 2개의 의사결정나무 모형만으로도 해석이 가능하므

로 의사결정나무 모형 생성 및 해석에 있어 효율적이라고 할 수 있다. 추후 연구 과제로 변수들 간의 관계 및 변수 축소에 있어 꼭 연관성 규칙 및 군집분석을 사용해야 하는 것은 아니므로 변수들 간의 관계 파악 및 다양한 변수들을 축소하여 새로운 변수로 추출하는 방법에 대하여 여러 가지 다양한 분석 방법을 접목해 볼 필요성이 있다.

참고문헌

[1] 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, pp.671-678. 2011.

[2] Cho, K. H. and Park, H. C. “A study on association rule creation by marginally conditional variables”, Journal of the Korean Data & Information Science Society, 23, pp.121-129. 2012.

[3] Cho, K. H. and Park, H. C. “A study on decision tree creation using marginally conditional variables”, Journal of the Korean Data & Information Science Society, 23, pp.299-307. 2012.

[4] Park, H. C. “Proposition of negatively pure association rule threshold”, Journal of the Korean Data & Information Science Society, 22, pp.179-188. 2011.

[5] Agrawal, R., Imielinski, R. and Swami, A. “Mining association rules between sets of items in large databases“, Proceedings of the ACM SIGMOD Conf. on Management of Data, pp.207-216. 1993.

[6] MacQueen, J. “Some methods for classification and analysis of multivariate observations”, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1, pp.281-297. 1967.