Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 4, pp. 927 ~ 935

A Study of Association Rule Mining by Clustering through Data Fusion

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

Abstract

Currently, Gyeongnam province is executing the social index survey every year to the provincials. But, this survey has the limit of the analysis as execution of the different survey per 3 year cycles. The solution of this problem is data fusion. Data fusion is the process of combining multiple data in order to provide information of tactical value to the user. But, data fusion doesn't mean the ultimate result. Therefore, efficient analysis for the data fusion is also important. In this study, we present data fusion method of statistical survey data. Also, we suggest application methodology of association rule mining by clustering through data fusion of statistical survey data.

Keywords : Association Rule, Clustering, Data Fusion, Hybrid Data Mining, Statistical Matching

1. 서론

현대 사회에서는 조직의 운영 및 의사 결정을 위하여 다양한 통계 조사가 실시되고 있으며, 연구의 목적에 따라 조사 문항을 다르게 하여 실시하고 있다. 이는 동일한 모집단에 대한 조사라 할지라도 조사의 문항이 다른 경우 각각의 개별적인 분석만 가능하며, 다른 조사 자료를 활용한다든지 이전 조사 자료와 연계한 분석이 가능하지 못하여 자료를 효율적으로 사용하고 있지 못하는 실정이다. 특히 경상남도의 경우 매년 사회지표 조사를 실시하고 있다.

사회지표 조사는 주민들이 생각하는 사회 상태를 총체적이고도 집약적으로 나타내는 것으로, 변화하는 역사적 흐름 속에서 우리가 처해 있는 사회적 상태를 종합적이

1) a part-time lecturer, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : cho1023@changwon.ac.kr

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

고 집약적으로 나타냄으로써 사회구성원들의 삶의 질을 전반적으로 파악하고 사회변화를 포착할 수 있는 척도이다. 현재 경상남도는 도민들을 대상으로 3년 주기로 매년 설문 문항을 다르게 하여 사회지표 조사를 실시하고 있어 도민들의 환경의식에 대한 분석 시, 연도별로 각각 분석을 실시해야 함으로써 유기적인 분석이 가능하지 못한 실정이다. 또한, 특정 연도의 사회지표 조사 자료에서는 환경의식 분석에 사용할 환경 관련 문항들이 기타 연도에 비하여 작아 다양한 분석을 실시하지 못하고 있다. 이에 각 연도의 사회지표 조사 자료를 결합하여 하나의 데이터 파일을 만들면 고부가가치의 정보를 획득할 수 있을 것이며, 이를 위하여 사용되어 지는 방법 중의 하나가 데이터 퓨전(data fusion) 방법이다.

데이터 퓨전은 같은 모집단에서 나온 서로 다른 표본들을 포함하는 데이터 셋을 합치는 기법 또는 처리과정으로 정의되며, 데이터 융합, 데이터 결합, 데이터 매칭이라고 불리기도 한다. 데이터 퓨전은 통계 분석의 최종 결과라기보다는 통계 분석 결과의 질을 높이기 위한 방법이라고 할 수 있다. 다시 말해서 데이터 퓨전을 통해서 얻은 최종 결과에 대한 추가된 정보를 이용함으로써 통계 분석의 질을 향상시킬 수 있는 방법이므로, 데이터 퓨전에 의해서 얻어진 정보를 효율적으로 분석하는 것 또한 중요하다. 이에 본 논문에서는 통계 조사 자료에 대하여 데이터 퓨전을 실시하고, 데이터 퓨전에 의해 생성된 자료에 대하여 하이브리드 데이터 마이닝(hybrid data mining) 방법인 집단 세분화에 의한 연관성 규칙을 적용하는 방안에 대하여 연구하고자 한다.

일반적으로 데이터 마이닝(data mining)은 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정으로, 대용량의 관측 가능한 데이터를 기반으로 숨겨진 지식, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 바탕으로 의사결정 등을 위한 정보로 활용하고자 하는 것이다. 데이터 마이닝 기법으로는 군집분석(cluster analysis), 연관성규칙(association rule), 의사결정나무기법(decision tree), 신경망모형(neural network), 자기조직화지도(self- organizing map; SOM) 등의 분석기법이 있으며, 데이터 마이닝은 이들 각각에 대한 하나의 기법만을 사용하여 분석을 실시한다. 반면, 하이브리드 데이터 마이닝은 데이터 마이닝이 수행하는 작업과 목적, 분석에 이용되는 데이터의 특성, 발견된 패턴의 설명력, 사용의 용이성 등에 따라 몇 개의 데이터 마이닝 기법을 결합함으로서 하나의 기법이 가지는 한계를 극복할 수 있어 효율적으로 데이터 마이닝을 수행 할 수 있게 한다. 이에 본 논문에서 제안하는 통계 조사 자료의 데이터 퓨전에 의한 집단 세분화에 의한 연관성 규칙의 적용 방안은 통계 조사 자료의 질을 높임과 동시에 효과적으로 데이터 마이닝 기법을 적용할 수 있어, 통계 조사 자료를 더욱더 총체적이고 효율적으로 분석할 수 있게 한다.

하이브리드 데이터 마이닝의 선행 연구로 강문식과 이상용(2002)은 비감독 학습 방법의 경쟁 학습 모델에서 정규화 과정의 위험 없이 목적 패턴을 도출하고 감독 학습 방법의 역전파 알고리즘에서 목적 패턴을 인위적으로 생성하지 않아 신경망의 큰 문제점인 정규화의 위험과 오버피팅(overfitting)을 방지할 수 있는 경쟁 학습 모델과 BP알고리즘을 결합한 하이브리드형 신경망 모델인 HACAB(Hybrid Algorithm Combining a competition leaning model And Bp algorithm)에 대하여 연구하였다. 김만선과 이상용(2002)은 인공지능적 기법인 자기 조직화 지도와 통계적 기법인 계층적 군집화 기법을 접목하는 방법으로 두 단계의 군집화를 수행하여 군집화를 수행하는 과정에서 소군집의 특징을 나타내는 요약 정보인 특정값을 이용하기 때문에 계산 속

도를 향상시킬 수 있는 PPC(Per Post Clustering)에 대하여 연구하였다. 윤경배 등(2002)은 KDD(Knowledge Discovery in Database) 분야에서 의사결정 관리에 필요한 지식을 얻기 위해 효율적으로 지식 베이스를 검색하고 갱신하는 관리 방법을 위하여 자율학습 신경망인 자기자기 조직화 지도에 확률적 분포 이론을 결합한 하이브리드 SOM을 제안하였다. 황인수(2002)는 군집분석 시 최적화된 그룹 세분화를 위하여 2단계 계층적 클러스터링 기법에 대하여 연구하였으며 김진성(2003)은 연관규칙과 퍼지 인공 신경망을 결합한 하이브리드 데이터 마이닝에 대하여 연구한 바 있다.

본 논문의 2장에서는 데이터퓨전 기반 집단 세분화에 의한 연관성 규칙의 배경에 관하여 기술하고, 3장에서는 데이터퓨전 기반 집단 세분화에 의한 연관성 규칙의 적용 방안에 대하여 기술한다. 4장에서는 데이터퓨전 기반 집단 세분화에 의한 연관성 규칙 적용 사례를 제시 한 후 5장에서 결론을 맺고자 한다.

2. 데이터퓨전 기반 집단 세분화에 의한 연관성 규칙의 배경

본 논문에서는 다양한 통계 조사를 하나의 파일로 결합하고, 이 결합된 파일을 바탕으로 집단 세분화를 실시한 후, 이 집단 세분화의 결과를 목표변수로 지정하여 최종적으로 의사결정나무 분석을 실시하는 방안을 연구하고자 한다. 여기서 다양한 통계 조사를 하나의 파일로 결합하기 위하여 데이터 퓨전 기법을 사용하였고, 집단 세분화를 도출하기 위하여 군집 분석을 사용하였다.

본 절에서는 본 논문에서 사용한 데이터 퓨전, 군집 분석, 의사결정나무에 대하여 기술하고자 한다. 데이터 퓨전은 같은 모집단에서 나온 서로 다른 표본들을 포함하는 데이터 셋을 합치는 기법 또는 처리 과정으로 정의된다.

데이터 퓨전은 별개의 데이터 파일을 결합하여 하나의 완전한 데이터 파일을 만드는 것을 의미하는 것으로 데이터 융합, 데이터 결합, 데이터 매칭이라고 불리기도 한다(한상훈 등, 2004). 데이터 퓨전은 그 자체가 하나의 분석이며 최종 결과라기보다는 통계분석 결과의 질을 높이기 위한 방법이라고 할 수 있다. 즉, 데이터 퓨전을 통해서 얻은 최종 결과에 대한 추가된 정보를 이용함으로써 통계 분석의 질을 향상시킬 수 있다. 영국 National Statistics(2003)에 따르면 데이터 퓨전의 종류는 정확 결합(exact matching), 판단 결합(judgemental matching), 확률적 결합(probability matching), 통계적 결합(statistical matching), 데이터 연결(data linking)로 구분된다.

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 제곱거리는 다음과 같다.

$$d_{ij}^2 = (X_i - X_j)'(X_i - X_j) \tag{2.1}$$

Agrawal 등(1993)에 의해 처음 소개된 연관성 규칙은 탐색적이며, 비목적성 분석이며, 기존의 데이터를 특별한 변형 없이 계산이 용이하게 사용 가능하다는 장점을 가지고 있으며, 계산 과정이 길고, 반복된 계산이 많으며, 적절한 품목의 결정이 어렵고, 각 품목의 단위에 따른 표준화가 어렵다는 단점을 아울러 가지고 있다. 연관성 규칙은 이러한 단점에도 불구하고 두 품목간의 관계를 명확히 수치화함으로써 두 개 이상의 품목간의 관련성을 표시하여 주기 때문에 현업에서 많이 활용되고 있다. 연관규칙을 평가하는 기준에는 지지도(support), 신뢰도(confidence), 향상도(lift) 등이 있으며 다음과 같다.

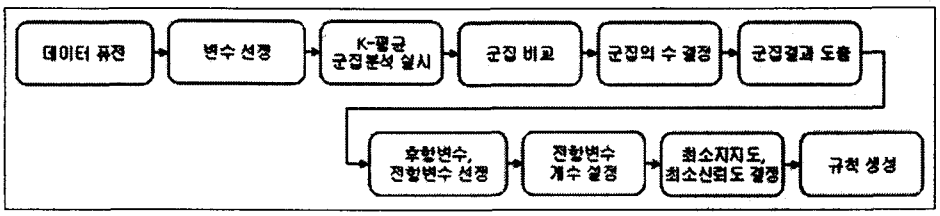
$$\text{지지도 : } S_{(X \Rightarrow Y)} = P(X \cap Y) \tag{2.2}$$

$$\text{신뢰도 : } C_{(X \Rightarrow Y)} = P(Y|X) = \frac{P(X \cap Y)}{P(X)} \tag{2.3}$$

$$\text{향상도 : } L_{(X \Rightarrow Y)} = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)} \tag{2.4}$$

3. 데이터퓨전 기반 집단 세분화에 의한 연관성 규칙의 적용 방안

일반적으로 군집분석은 관찰 대상을 분류하거나 분류된 관찰 대상의 속성을 파악하는데 이용되어 군집분석만으로 분석이 종료되게 된다. 만일 군집분석의 결과 군집이 확연하게 구분되는 경우, 군집 분석의 결과를 바탕으로 이를 연관성 규칙에 적용하면 각 군집과 기타 조사 항목에 대한 관련성 여부를 분석하는데 효율적으로 사용할 수 있을 것이다. 이에 본 절에서는 데이터퓨전 기반 집단세분화에 의한 연관성 규칙 적용 방안을 제시하고자 한다. 데이터퓨전 기반 집단세분화에 의한 연관성 규칙 적용 단계는 <그림 2>와 같다.



<그림 2> 집단세분화에 의한 연관성 규칙 적용 단계

[단계 1] 데이터 퓨전

본 논문에서는 데이터 퓨전 알고리즘 중 통계적 결합 방법을 사용하였다. 통계적 결합은 공통으로 가지는 변수는 존재하나 고유 식별 변수처럼 개인 식별 가능한 변수가 없을 때 회귀분석, 로지스틱 회귀분석 등을 사용하여 통계적 방법을 사용하여 데

이터 결합하는 방법이다(한상훈 등, 2004). 현재 데이터 퓨전을 위한 소프트웨어는 개발되어 있지 않으며, 본 논문에서는 통계적 결합에 대한 SAS 매크로 프로그램을 사용하여 데이터 퓨전을 실시한다(박희창과 조광현, 2006; 조광현과 박희창, 2007).

[단계 2] 변수 선정
k-평균 군집분석에 사용할 변수를 선정한다.

[단계 3] k-평균 군집분석 실시
군집의 특성이 명확하게 파악되는 군집을 도출하기 위하여 2에서 5개의 군집으로 k-평균 군집분석을 실행한다.

[단계 4] 군집 비교
2에서 5개의 군집으로 k-평균 군집분석을 실시한 후 각각의 군집의 비교한다.

[단계 5] 군집의 수 결정
군집의 결과가 가장 확연하게 구분되는 군집의 수를 결정한다.

[단계 6] 군집결과 도출
연관성 규칙에 적용하기 위하여 군집 분석의 결과를 변수로 저장한다.

[단계 7] 변수선정
연관성 규칙을 생성하기 위하여 전항변수와 후항변수를 선정한다. 군집 분석의 결과에 의한 변수를 후항변수로 선정하고, 인구통계학 속성 관련문항을 전항변수로 선정한다.

[단계 8] 전항변수 개수 결정
변수들 간의 관련성을 알아보기 위하여 전항변수의 개수를 결정해야 한다. 전항변수의 개수가 적으면 연관성 규칙 모형의 정확도는 증가하나 의미 있는 규칙을 찾아내지 못할 수 있고, 전항변수의 개수가 많으면 많은 규칙을 생성하나 연관성 규칙 모형의 정확도가 감소할 수 있다.

[단계 9] 최소지지도, 최소신뢰도 결정
연관성 규칙 생성에 있어 최소지지도와 신뢰도를 결정해야 한다. 최소지지도와 신뢰도를 낮게 결정하면 연관성 규칙의 생성이 많아지나 의미 없는 규칙이 생성될 수 있고, 최소지지도와 신뢰도를 높게 하면 의미 있는 규칙을 찾아내지 못하게 되는 경우도 있다.

[단계 10] 규칙생성
모형 평가에서 모형이 적합하다고 판단되면 환경관련 문항 각각에 대하여 연관성 규칙을 생성한다.

4. 데이터퓨전 기반 집단 세분화에 의한 연관성 규칙 적용 사례

4.1 데이터 퓨전

본 절에서는 2001년 조사된 사회지표 조사와 2002년 조사된 사회지표 조사 및 2003년 조사된 사회지표 조사의 환경자료에 대하여 데이터 퓨전 기법을 적용한다. 2001년과 2002년 및 2003년에 조사된 사회지표 조사 자료에서는 2001년과 2002년 자료의 데이터 퓨전, 2001년과 2003년 자료의 데이터 퓨전, 2002년과 2003년 자료의 데이터 퓨전 등 총 3번의 데이터 퓨전을 실시 한 후, 각각의 데이터 퓨전 결과를 통합하여 최종 퓨전된 데이터 파일을 생성한다.

각각 데이터는 약 10,000건이며, 2001년 사회지표 조사에서는 환경관련 6문항과 인구통계학적 속성 6문항으로 구성되어 있고, 2002년 사회지표 조사에서는 환경관련 3문항과 인구통계학적 속성 6문항으로 구성되어 있으며, 2003년 사회지표 조사에서는 환경관련 2문항과 인구통계학적 속성 6문항만으로 구성되어 있다.

데이터 퓨전에 의한 최종 자료는 <표 1>과 같다. <표 1>에서 보는 바와 같이 데이터 퓨전에 의한 최종 자료의 레코드 수는 총 29,876건이다. 변수는 인구통계학적 속성 문항인 연령, 주관적 사회계층, 학력, 성별, 결혼유무, 거주지역의 6개 문항과 2001년의 환경관련 6문항, 2002년의 환경관련 3문항, 2003년 환경관련 2문항 등 총 17개 문항으로 구성되어 있다. <표 1>에서 연속형 자료는 등간 및 비율자료를 의미하고, 범주형 자료는 명목 및 서열자료를 의미하며, 환경관련문항에서의 연속형 자료와 인구통계학적 속성문항에서의 주관적 사회계층 문항은 편의상 등간자료로 하여 분석하고자 한다.

<표 1> 데이터 퓨전에 의한 최종 자료

변수 구분	변수명	자료 구분
환경관련문항	지역의 상수도 환경오염도	연속형자료
	지역의 하수도 환경오염도	연속형자료
	지역의 소음진동 환경오염도	연속형자료
	지역의 악취 환경오염도	연속형자료
	지역의 대기 환경오염도	연속형자료
	지역의 토양 환경오염도	연속형자료
	쓰레기 분리수거의 참여 정도	연속형자료
	녹색 제품의 구입 여부	범주형자료
	수돗물 음용수 적정 여부	연속형자료
	자녀의 환경오염 저감 행동 유무	범주형자료
	산림의 중요성 체감도	연속형자료
인구통계학적 속성 문항	연령	연속형자료
	주관적 사회계층	연속형자료
	학력	범주형자료
	성별	범주형자료
	결혼유무	범주형자료
	거주지역	범주형자료
총 데이터 건수	29,876 건	

4.2 데이터 퓨전 기반 집단 세분화에 의한 연관성 규칙

데이터 퓨전 기반 집단세분화에 의한 연관성 규칙에서는 군집분석에 의한 결과를 도출하기 위하여 <표 1>의 결과 중 지역의 상수도 환경오염도, 지역의 하수도 환경오염도, 지역의 소음진동 환경오염도, 지역의 악취 환경오염도, 지역의 대기 환경오염도, 지역의 토양 환경오염도의 총 6개 환경관련 문항에 대하여 군집분석을 실시하였다. k-평균 군집분석 시, 군집의 특성이 명확하게 파악되는 군집을 도출하기 위하여 2에서 5개의 군집으로 군집분석을 실행한 결과 2개의 군집으로 나누었을 때 군집의 특성이 명확하게 구분되었다. 군집의 수를 2개로 하여 k-평균 군집분석을 실시한 결과는 <표 2>와 같다.

<표 2>에서 보는 바와 같이 1 군집의 집단은 지역의 환경 오염도에 대한 응답이 긍정적인 집단으로 분류되고, 2 군집의 집단은 지역의 환경 오염도에 대한 응답이 부정적인 집단으로 분류되었다. k-평균 군집분석에 의한 결과를 후항변수로 지정하여 연관성 규칙을 적용하였다. 연관성 규칙 적용 시, 전항 변수의 개수를 1로 지정하고 최소 지지도를 10으로 최소 신뢰도를 80으로 지정하여 분석을 실시하였다.

<표 2> k-평균 군집분석 결과(2군집)

항목 \ 군집	1 군집	2 군집
지역의 상수도 환경오염도	3.509	2.924
지역의 하수도 환경오염도	3.658	2.890
지역의 소음진동 환경오염도	3.920	2.891
지역의 악취 환경오염도	4.122	3.210
지역의 대기 환경오염도	4.001	3.028
지역의 토양 환경오염도	3.928	3.004
군집 레코드 수	18,154	11,822

데이터 퓨전 기반 집단세분화에 의한 연관성 규칙 생성 시, 생성된 모형의 예측 정확도가 기존의 결과변수들에 대한 모형 예측 정확도보다 현저히 떨어진다면 집단세분화에 의한 연관성 규칙은 의미가 없을 것이다. 여기서 모형의 예측 정확도는 연관성 분석에 의해 생성된 규칙이 얼마나 정확한가를 알아보는 모형의 평가 기준으로서 모형의 예측 정확도는 "1-(오분류율)"을 의미한다.

모형의 정확도를 비교하기 위하여 자료를 훈련 자료와 모형평가 자료로 분할한 뒤, 기존의 후항변수에 대한 모형의 예측 정확도 및 모형평가 예측 정확도와 집단세분화에 의한 모형의 예측 정확도 및 모형평가 예측 정확도를 비교하였다. 여기서 훈련 자료와 모형평가 자료를 각각 1/2로 나누었으며, 훈련 자료는 연관성 규칙 모형을 생성하는데 이용하고 모형 평가 자료는 생성된 모형이 얼마나 정확한가를 확인하기 위하여 사용한다. 즉, 원 자료를 두 개의 자료로 분할하고 하나는 모형의 생성을 위하여 사용하고, 나머지 하나는 생성된 모형의 평가를 위하여 사용한다는 의미이다.

기존의 후항변수에 대한 모형 예측 정확도 및 모형평가 예측 정확도와 집단세분화 변수에 대한 모형 예측 정확도 및 모형평가 예측 정확도 비교는 <표 3>과 같다. 여기서 모형 예측 정확도는 생성된 연관성 규칙 모형 자체에 대한 예측 정확도를 의미

하며, 모형평가 예측 정확도는 생성된 모형을 모형평가 자료에 적합하여 계산되어진 예측 정확도를 의미한다.

<표 3>에서 보는 바와 같이 집단 세분화 변수인 지역의 환경오염도에 대한 모형 예측 정확도와 모형평가 예측 정확도가 원 변수인 지역의 토양 환경오염도의 예측 정확도보다는 다소 낮게 나타났으나, 지역의 상수도 환경오염도, 지역의 하수도 환경오염도, 지역의 소음진동 환경오염도, 지역의 악취 환경오염도, 지역의 대기 환경오염도에 비해서는 예측 정확도가 높게 나타나 집단 세분화에 의한 연관성규칙이 효율적임을 알 수 있다.

<표 3> 모형 예측 정확도 비교

결과변수 \ 정확도		모형 예측 정확도	모형평가 예측 정확도
원 변수	지역의 상수도 환경오염도	74%	71%
	지역의 하수도 환경오염도	76%	73%
	지역의 소음진동 환경오염도	78%	75%
	지역의 악취 환경오염도	76%	76%
	지역의 대기 환경오염도	78%	75%
	지역의 토양 환경오염도	82%	81%
집단 세분화 변수	지역의 환경오염도	81%	80%

지역의 환경 오염도에 대한 연관성 규칙 결과는 <표 4>와 같다.

<표 4> 지역의 환경 오염도에 대한 연관성 규칙 결과

규칙	지지도	신뢰도	후항변수	전항변수
1	32.4	85.4	지역의 환경오염도 = 불만	성별 = 여성
2	28.7	83.1	지역의 환경오염도 = 불만	연령 = 평균이하
3	42.9	81.9	지역의 환경오염도 = 불만	주거지역 = 주거, 상가, 공업지역
4	35.4	81.0	지역의 환경오염도 = 불만	주관적 사회계층 = 중산층 미만
5	29.5	80.3	지역의 환경오염도 = 불만	학력 = 고졸이하

<표 4>에서와 같이 지역의 환경 오염도에 대한 연관성 규칙 결과를 자세히 살펴보면, 성별이 여성, 연령이 평균이하, 주거지역이 주거, 상가, 공업 지역, 주관적 사회계층이 중산층 미만, 학력이 고졸이하인 응답자들은 지역의 환경 오염도에 대하여 불만의 응답 높은 것으로 나타났다.

5. 결론

현재 경상남도는 경상남도 도민들을 대상으로 매년 환경, 교통 등의 부문에 대하여 사회지표 조사를 실시하고 있다. 그러나 3년 주기로 매년 설문 문항을 다르게 하여 설문조사를 실시하고 있어 도민들의 환경의식에 대한 분석 시 연도별로 각각 분석을 실시해야 함으로서 유기적인 분석이 가능하지 못하도록 하는 원인이 되어, 결국에는 고부가가치의 정보를 얻는데 어려움을 주게 된다. 그러므로 본 논문에서 제시한 데이

터 퓨전으로 각 통계 조사 자료를 결합하여 하나의 데이터 파일로 만들면 데이터의 질을 높일 수 있다. 그러나 데이터 퓨전은 최종 결과라기보다는 통계 분석 결과의 질을 높이기 위한 방법이라고 할 수 있으므로, 데이터 퓨전에 의한 효율적인 분석 방법의 적용 또한 중요한 과제이다. 이에 본 논문에서는 데이터 퓨전에 의해 생성된 자료를 기반으로 한 집단 세분화에 의한 연관성 규칙을 적용하는 방안에 대하여 연구하였다. 일반적으로 군집분석은 관찰 대상을 분류하거나 분류된 관찰 대상의 속성을 파악하는데 이용되어 군집분석만으로 분석이 종료되게 된다. 만일 군집분석의 결과 군집이 확연하게 구분되는 경우, 군집 분석의 결과를 바탕으로 이를 연관성 규칙에 적용하면 각 군집과 기타 조사 항목에 대한 관련성 여부를 분석하는데 효율적으로 사용할 수 있는 장점이 있으며, 이는 적용 사례를 통하여 확인할 수 있었다.

참고문헌

1. 강문식, 이상용 (2002). 데이터 마이닝을 위한 경쟁학습모델과 BP 알고리즘을 결합한 하이브리드 신경망, *Journal of information technology application& management*, Vol. 9, No. 2, pp.1-16.
2. 김만선, 이상용 (2003). 대용량 데이터 처리를 위한 하이브리드형 클러스터링 기법, *정보처리학회논문지*, Vol. 10-B, pp.33-40.
3. 김진성 (2003). 연관규칙과 퍼지 인공신경망에 기반한 하이브리드 데이터 마이닝 매커니즘에 관한 연구, *한국경영학회/대한산업공학회* 2003 춘계공동학술대회, pp.884-888.
4. 박희창, 조광현 (2006). 통계적 데이터 퓨전을 위한 sas 매크로, *한국자료분석학회논문지*, Vol. 8, No. 5, pp.1927-1937.
5. 윤경배, 최준혁, 왕창종(2002). 하이브리드 SOM을 이용한 효율적인 지식베이스 관리, *정보처리학회논문지*, Vol. 9-B, pp.635-642.
6. 조광현, 박희창 (2007). Association Rule Mining by Environmental Data Fusion, *한국데이터정보과학회지*, Vol. 18, No. 2, pp.279-287.
7. 한상훈, 안일호, 하덕주, 최종후 (2004). 데이터 퓨전과 평가, *한국데이터마이닝학회 2004 추계학술대회*, pp.238-254.
8. 황인수 (2002). 데이터 마이닝에서 그룹 세분화를 위한 2단계 계층적 클러스터링 알고리듬, *한국경영학회지*, Vol. 19, No. 1, pp.189-196.
9. Agrawal R., Imielinski R., Swami A. (1993), "Mining association rules between sets of items in large databases", In *Proc. of the ACM SIGMOD Conference on Management of Data*, Washington, D. C.
10. National Statistics (2003), National Statistics Code of Practice Protocol on Data Matching. http://www.statistics.gov.uk/about/consultations/general_consultations /downloads/ Protocol_on_Data_Matching.pdf

[2007년 9월 접수, 2007년 11월 채택]