응용통계연구 제 **21**권 *3*호, *2008*년, *pp. 399-411*

통계조사에서의 퓨전된 자료에 대한 하이브리드 데이터마이닝의 적용 방안

박희창[1] 조광현[2]

요 약

현대 사회에서는 조직의 운영 및 의사 결정을 위하여 다양한 통계 조사가 실시되고 있으며, 연구의 목적에 따라 조사 문항을 다르게 하여 실시하고 있다. 현재 경상남도의 경우 3년 주기로 매년 설문 문항을 다르게 하여 사회지표 조사를 실시하고 있어 유기적인 분석이 가능하지 못한 실정이다. 이에 본 장에서는 데이터 퓨전을 이용하여 다양한 통계 조사 자료를 결합하여 고부가적인 자료를 생성하고자 한다. 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 데이터마이닝 기법인 잠재변수를 이용한 신경망 분석을 적용하는 방안에 대하여 연구하고자 한다.

주요용어: 데이터 퓨전, 신경망 분석, 잠재변수, 하이브리드 데이터마이닝.

1. 서론

현대 사회에서는 조직의 운영 및 의사 결정을 위하여 다양한 통계 조사가 실시되고 있다. 현재 통계 조사는 연구의 목적에 따라 조사 문항을 다르게 하여 실시하고 있다. 이는 통계 분석 시, 통계 조사의 개별적인 분석만 가능하여 통계 조사 자료의 효율적 사용이 제한되어 있어, 동일한 모집단에 대한 조사라 할지라도 조사의 문항이 다른 경우 각각의 개별적인 분석만 가능하다는 점과 다른 조사 자료를 활용한다든지 이전 조사 자료와 연계한 분석이 가능하지 못하다는 문제점이 있다. 이에, 통계 조사 자료를 효율적으로 사용하기 위하여 서로 다른 통계 조사 자료를 하나의 데이터 파일로 만들면 고부가가치의 정보를 획득할 수 있으며, 통계 조사 간의 유기적인 분석이 가능하다.

예를 들어, 경상남도에서 조사되고 있는 사회지표조사의 경우, 조사 항목이 많아 3년 주기로 매년 설문 문항을 다르게 하여 설문조사를 실시하고 있어 환경 관련 분석 시 2001년과 2002년 및 2003년 조사된 사회지표 조사의 환경자료를 각각 분석해야 하며, 2001년과 2002년 및 2003년 조사된 사회지표조사의 환경 관련 문항을 통합적으로 분석을 할 수 없는 문제점

1) (641-773) 경상남도 창원시 사림동 9번지, 창원대학교 통계학과, 교수.
교신저자: hcpark@changwon.ac.kr
2) (641-773) 경상남도 창원시 사림동 9번지, 창원대학교, 시간강사.
E-mail: cho1023@changwon.ac.kr

이 있다. 그러나 데이터 퓨전에 의한 2001년과 2002년 및 2003년의 사회지표조사를 결합하면 환경관련 문항의 응답 결과를 종합적으로 분석할 수 있다. 이에 본 논문에서는 2001년과 2002년 및 2003년의 사회지표조사에 대한 데이터 퓨전 기법의 적용으로 다양한 통계 조사 자료를 결합하여 고부가적인 자료를 생성하고자 한다. 여기서 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 대한 하이브리드 데이터 마이닝 방법인 잠재변수(latent variable)를 이용한 신경망 분석을 적용하는 방안에 대하여 연구하고자 한다.

일반적으로 데이터 마이닝(data mining)은 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정으로 군집 분석(cluster analysis), 연관성규칙(association rule), 의사결정나무기법(decision tree), 신경망모형(neural network), 자기조직화지도(self-organizing map: SOM) 등의 분석 기법이 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 조사에서의 퓨전된 자료에 대한 하이브리드 데이터 마이닝 방법인 잠재변수를 이용한 신경망 분석의 적용 방안은 통계 조사 자료의 질을 높임과 동시에 효과적으로 데이터 마이닝의 적용을 가능하게 할 수 있다.

하이브리드 데이터 마이닝의 선행 연구로 강문식과 이상용 (2002)은 경쟁 학습 모델과 BP알고리즘을 결합한 하이브리드형 신경망 모델인 HACAB(Hybrid Algorithm Combining a competition leaning model And Bp algorithm)에 대하여 연구하였다. 김만선과 이상용 (2003)은 인공지능적 기법인 자기 조직화 지도와 통계적 기법인 계층적 군집화 기법을 접목하는 방법인 PPC(Per Post Clustering)에 대하여 연구하였다. 윤경배 등 (2002)은 KDD(Knowledge Discovery in Database) 분야에서 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 통계조사에서의 퓨전된 자료에 대한 잠재변수를 이용한 신경망 분석의 이론적 배경에 관하여 기술하고, 3장에서는 적용 방안에 대하여 기술한다. 4장에서는 적용 예제를 제시 한 후 5장에서 결론을 맺고자 한다.

2. 이론적 배경

본 논문에서는 데이터 퓨전에 의해 생성된 자료에 대한 하이브리드 데이터 마이닝 방법인

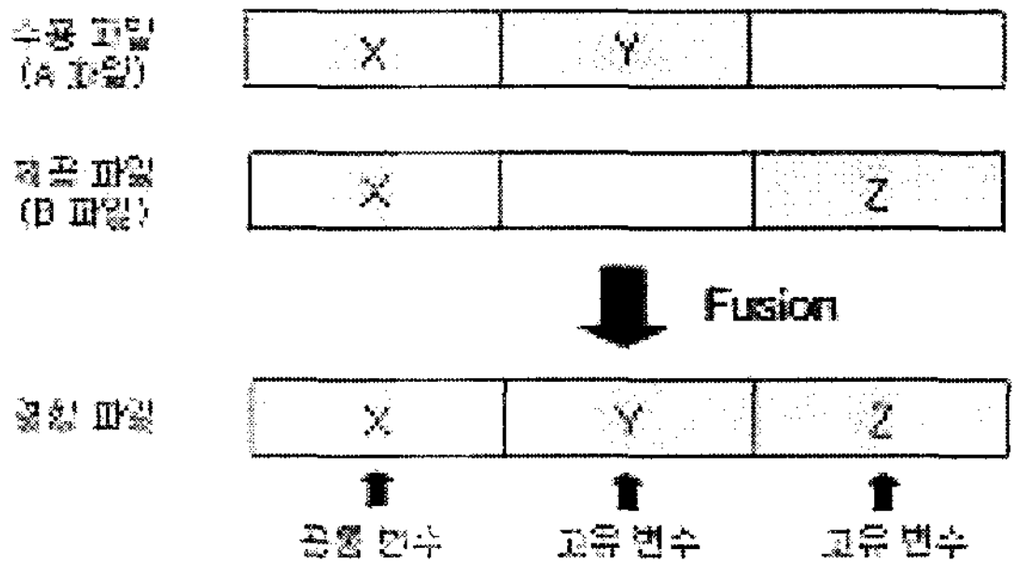

그림 2.1: 데이터 퓨전

잠재변수를 이용한 연관성규칙을 적용하는 방안을 제안하고자 한다. 본 절에서는 데이터 퓨전, 잠재변수, 신경망 분석에 대하여 기술하고자 한다.

첫 번째로 다양한 통계 조사에 대한 데이터 퓨전을 실시한다. 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.

서로 다른 두 개의 파일 A와 B를 가정하자. 파일 A와 B에는 X라는 변수가 공통적으로 존재하고 Y변수와 Z변수는 A 파일과 B 파일에 각각 존재한다고 하자. 즉, 파일 A와 B에 공통적으로 X라는 변수가 있고 파일 A에는 Y라는 변수만 존재하며 파일 B에는 Z라는 변수만 존재한다고 하자. 변수 X, Y, Z로 구성된 파일을 만들기 위하여 그림 2.1과 같이 파일 A와 B를 결합하여 하나의 파일로 만들면 된다.

파일 A에는 변수 X와 변수 Y로 구성되어 있고 파일 B에는 변수 X와 변수 Z로 구성되어 있다. 여기서 파일 A와 파일 B에 공통적으로 존재하는 변수 X를 공통 변수(common variable)라고 하고 파일 A 또는 파일 B에서만 존재하는 변수 Y와 변수 Z를 고유 변수(unique variable)라고 한다. 데이터 퓨전의 결과로 생성된 결합 파일은 두 파일의 공통변수 X를 이용하여 파일 B에 존재하는 변수 Z를 파일 A에 추가한 형식으로 나타난다. 여기서 변수 Z를 수용하는 파일 A를 수용 파일(recipient file)이라 하고 변수 Z를 제공하는 파일 B를 제공 파일(*donor file*)이라고 한다. 파일 A와 B에 의하여 데이터 퓨전을 수행한 후 생성된 파일을 결합 파일(fused file)이라고 한다. 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.

두 번째로 데이터 퓨전으로 생성된 자료를 바탕으로 잠재변수를 도출한다. 잠재변수란 관찰된 여러 변수들 중에서 서로 연관성이 있는 변수들끼리 묶여지는 변수를 의미이며, 주성분분석의 주성분(principal component), 요인분석의 요인(factor), 구조방정식의 구

조(constract)와 동일한 의미이다. 일반적으로 잠재 변수를 도출하기 위하여 요인분석을 가장 많이 사용한다. 요인분석은 다수 변수들 간의 상관관계를 이용하여 변수들 간의 체계적인 구조를 밝히고 서로 유사한 변수끼리 묶어주는 다변량 통계기법 중의 하나이다. 즉, 여러 개의 변수들이 서로 어떻게 연결되어 있는가를 분석하여 이들 변수간의 관계를 공동요인을 이용하여 설명하는 분석 기법이다. 요인분석을 이용하여 변수의 형태로 주어진 많은 정보를 쉽고 간단하게 적은 수의 요인으로 제시해 준다. 따라서 요인분석은 과다한 정보로 인한 문제를 해결해 주고, 자료의 성격을 쉽게 파악할 수 있도록 도와준다.

요인분석에서 $X' = (X_1, X_2, \ldots, X_n)$가 평균 μ, 공분산행렬 Σ를 갖는 다변량 정규분포를 갖는다고 할 때, 각 변수는 공통요인(common factor)과 특정요인(specific factor)으로 나눌 수 있다. 이를 일반적인 모형으로 나타내면 수식 (2.1)과 같다

$$X - \mu_i = l_{i1}F_1 + l_{i2}F_2 + \cdots + l_{im}F_m + \varepsilon_i, \quad i = 1, 2, \ldots, p, \tag{2.1}$$

여기서 F_j는 j번째 요인을 나타나고 ε_i는 i번째 특정요인(specific factor)을 나타내며, j번째 요인에서 i번째 변수의 요인적재량 l_{ij}는 변수 X_i와 요인 F_j의 공분산으로 계산되어진다.

마지막으로 도출된 잠재변수를 이용하여 최종적으로 신경망 분석을 적용한다. 신경망분석은 신경생리학부에서 두뇌의 활동을 이해하고자 하는 목적 하에 신경의 작업을 설명하려는 시도에서 출발하여, 생물학적인 프로세스를 컴퓨터를 이용하여 모형화하는 분석 방법이다. 신경망분석은 1980년대 이후 생물학적 활동의 모형발전과 더불어 컴퓨터 성능의 진보, 신경망이론에 대한 통계학적인 접목으로 인해 빠르게 진보되면서 최근에는 데이터 마이닝에 있어서 유용한 기법이 되고 있다. 일반적으로 신경망분석은 고객의 신용평가, 불량거래의 색출, 의료진단예측, 우량고객의 선정, 타겟마케팅 등을 비롯한 여러 분야에 적용될 수가 있는데, 주로 지시(supervised) 데이터에 적용되어 결과변수(target)에 대한 예측이나 분류를 목적으로 감춰진 패턴을 찾고 이를 일반화하는데 이용된다. 신경망분석의 대표적인 알고리즘으로는 델타학습규칙을 일반화시킨 역전파알고리즘(backpropagation algorithm: BP)을 가장 많이 사용한다. 역전파알고리즘은 임의로 주어진 연결강도를 이용하여 출력값을 계산하고 계산된 출력값과 목표값의 차이와 오차를 계산한 뒤 오차값을 은닉층과 입력층으로 역전파 시켜 가중치를 조절해감으로써 분석을 실시한다 (강현철 등, 1999). 역전파알고리즘은 오차의 제곱으로 나타나는 비용 함수를 최소화시키는 방법으로 가장 많이 사용되는 오차의 제곱 비용함수는 수식 (2.2)와 같다.

$$E = \frac{1}{2}\sum_{i=1}^{p} \|y_i - d_i\|^2, \tag{2.2}$$

여기서 y_i는 실제 출력값이고 d_i는 목표값이다.

3. 적용 방안

일반적으로 신경망분석에서는 관심의 대상이 되는 결과변수에 대하여 모형화를 실시한다. 신경망분석은 일반적으로 입력변수와 결과변수의 관계를 그리기가 어려운 복잡한 데이터에

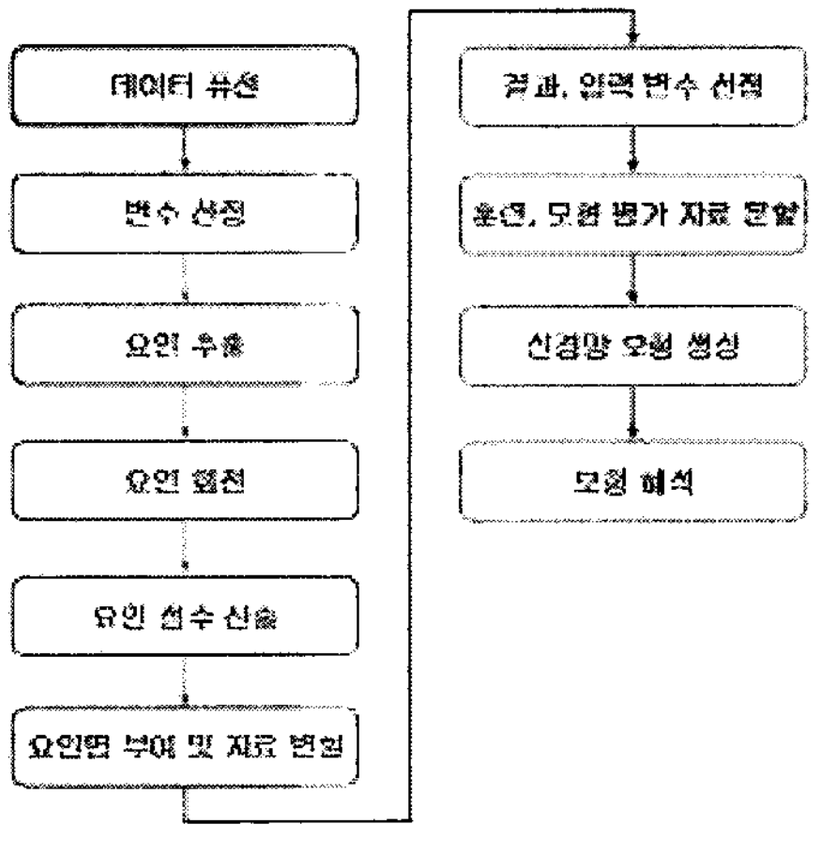

그림 3.1: 퓨전된 자료에 대한 잠재변수를 이용한 신경망 분석 절차

대해서도 좋은 결과를 주는 것으로 알려져 있으나, 모형화에 대한 시간이 많이 걸리고 생성된 모형에 대한 결과물을 제공할 뿐 어떠한 변수가 중요한지, 어떻게 상호작용이 이루어져 그러한 결과물을 주게 되는 지에 대한 설명은 하지 못하여 해석이 어렵다는 단점이 있다. 특히 여러 개의 결과변수를 대상으로 신경망 분석을 실시할 경우, 각각의 모형을 생성하고 이를 분석해야 하므로 많은 시간과 노력이 필요하다. 여러 개의 결과변수에 대한 유사성을 파악하여 몇 개의 변수로 축약할 수 있다면, 신경망의 모형 생성 및 모형 해석에 대한 시간을 단축할 수 있어 효율적인 신경망분석이 가능할 것이다. 이에 본 절에서는 데이터 퓨전 자료에 대하여 요인분석을 실시하여 여러 개의 결과변수에 대한 잠재 변수를 도출 한 후, 도출된 잠재 변수를 바탕으로 신경망분석을 실시하는 방안을 제시하고자 한다. 퓨전된 자료에 대한 잠재변수를 이용한 신경망분석의 절차는 그림 3.1과 같다.

[**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 데이터 결합하는 방법이다 (한상훈 등, 2004). 현재 데이터 퓨전을 위한 상용 소프트웨어는 개발되어 있지 않으며, 본 논문에서는 한상훈 등 (2004)에 의해 연구되어진 통계적 결합 방법을 사용하였다. 통계적 결합 알고리즘에서는 결합하고자 하는 변수가 범주형인 경우와 연속형인 경우로 구분되며, 결합변수는 제공파일에서만 존재하는 고유변수로서 수용파일에 결합하고자 하는 변수를 의미한다. 결합변수가 범주형인 경우에는 로지스틱 회귀분석을 이용하고, 결합변수가 연속형인 경우에는 회귀분석을 이용하여 데이터 퓨전을 실시하며, 본 논문에서는 박희창과 조광현 (2006)의 통계적 방법의 데이터 퓨전에 대한 SAS 매크로 프로그램을 사용하여 실제 자료에 데이터 퓨전을 실시한 조광현과 박희창 (2007)의 데이터 퓨전 결과를 이용한다.

[**단계 2**] 변수 선정: 잠재변수 도출을 위하여 요인분석에서 사용할 변수들을 선정한다.

[단계 3] 요인 추출: 요인 추출 방법에는 고유값(eigen value)을 기준으로 결정하는 방법, 총 분산 중 요인이 설명해 주는 정도를 기준으로 정하는 방법 그리고 연구자가 사전에 요인의 수를 결정하는 방법의 세 가지 방법이 있다. 이중 고유값을 기준으로 요인을 결정하는 방법이 가장 많이 사용된다.

[단계 4] 요인 회전: 요인 추출법에 의해 구해진 요인들이 뚜렷한 의미를 갖지 못하는 경우가 많이 있다. 변수들이 여러 요인에 비슷하게 요인 적재량을 나타낼 경우에는 변수들을 어떤 요인에 분류할 것인가를 결정하기가 쉽지 않다. 따라서 변수들을 어느 한 요인에 쏠리도록 요인을 회전시키면 해석이 쉽게 된다.

[단계 5] 요인점수 산출: 요인분석을 통해 얻어진 새로운 요인들은 차후의 분석에 이용하기 위하여 각 응답자별로 새로운 요인점수를 계산해 준다. 요인점수를 구하기 위한 방법으로는 가중된 최소 제곱 방법(weighted least square method)과 회귀분석 방법(regression method)이 있다.

[단계 6] 요인명 부여 및 자료변환: 각 요인들에 대한 요인명을 부여하고 산출된 요인 점수를 연관성 규칙에 사용하기 위하여 자료를 변환한다. 자료의 변환 시 이분형으로 변환을 실시한다.

[단계 7] 결과변수 및 입력변수 선정: 신경망분석의 모형화를 위하여 결과변수와 입력변수를 선정한다. 여기서 결과변수는 요인분석을 통하여 추출된 잠재변수를 사용한다.

[단계 8] 훈련 자료와 모형 평가 자료로 분할: 구축된 모형의 평가를 위하여 자료를 훈련 자료와 모형평가 자료로 분할한다. 모형 평가 시, 훈련 자료로 신경망분석의 모형을 구축하고 구축된 모형에 대하여 모형평가 자료를 적합하여 모형평가를 실시한다.

[단계 9] 신경망 모형 생성: 신경망 분석에 대한 모형 선택 기준(model selection criteria), 신경망 구조, 연결 강도 등을 지정하여 신경망 모형을 생성한다.

[단계 10] 모형 해석: 생성된 신경망 모형에 대한 해석을 실시한다.

4. 적용 예제

본 절에서는 2001년 조사된 사회지표 조사와 2002년 조사된 사회지표 조사 및 2003년 조사된 사회지표 조사의 환경자료에 대하여 데이터 퓨전 기법을 적용한다. 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 앞서 2001년과 2002년 및 2003년에 조사된 사회지표 조사에 대한 조사 대상 모집단들의 분포의 형태가 동일하지에 대한 동질성 검정을 실시한 결과, 동질성을 만족하는 것으로 나타났다. 2001년, 2002년, 2003년 조사된 사회지표 조사 자료에 대한 데이터 퓨전 방법은 그림 4.1과 같다. 데이터 퓨전에 의한 최종 자료의 레코드 수는 총

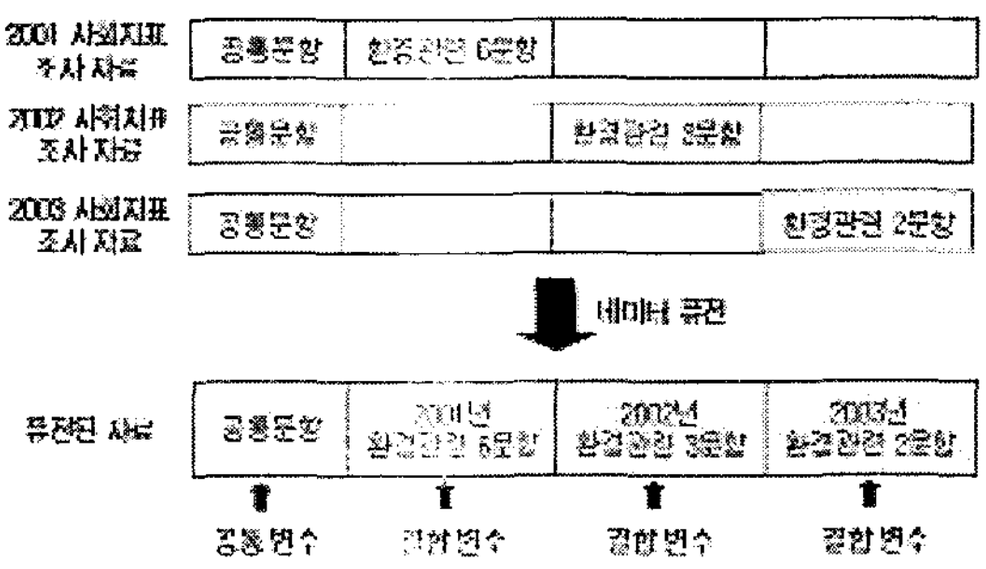

그림 4.1: 사회지표 조사의 데이터 퓨전

표 4.1: 잠재변수 추출 결과

요인	잠재변수(요인명)	변수
1 요인	주관적 수질 오염도	주관적 상수도 오염도, 주관적 하수도 오염도
2 요인	주관적 공간 오염도	주관적 소음진동 오염도, 주관적 악취 오염도 주관적 대기 오염도, 주관적 토양 오염도

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 (조광현과 박희창, 2007). 여기서 데이터 퓨전 시 3개의 데이터 셋을 동시에 퓨전하는 것이 아니라 데이터 셋에 공통으로 존재하는 변수를 이용하여 2001년 자료를 수용파일로 지정하고 2002년 자료를 제공파일로 지정하여 퓨전을 실시하였다. 또한 2001년 자료를 수용파일로 지정하고 2003년 자료를 제공파일로 지정하여 퓨전을 실시하는 등의 2개의 데이터 셋을 이용하여 여러 번의 데이터 퓨전을 실시한 결과를 통합하는 과정을 거쳤다. 또한 결측값의 처리는 평균값으로 대체를 하였다.

본 논문에서 제시하는 하이브리드 데이터 마이닝의 적용에 앞서 퓨전된 자료 중 관심 대상이 되는 주관적 상수도 오염도, 주관적 하수도 오염도, 주관적 소음진동 오염도, 주관적 악취 오염도, 주관적 대기 오염도, 주관적 토양 오염도의 총 6개 환경관련 문항과 인구통계학적 문항과의 예측 모형을 생성하기 위하여 신경망 분석을 실시하였다. 신경망 분석 결과, 몇 개의 변수들에 대하여 동일한 신경망 모형이 생성됨을 알 수 있었다. 이에 각 변수들에 대한 공통된 몇 개의 잠재변수를 추출하여 신경망 모형을 생성하다면 더욱 더 효과적으로 데이터 마이닝을 수행할 수 있을 것이라는 판단 하에 본 논문에서 제시하는 방법을 이용하여 분석을 실시하였다.

우선 잠재변수를 도출하기 위하여 주관적 상수도 오염도, 주관적 하수도 오염도, 주관적 소음진동 오염도, 주관적 악취 오염도, 주관적 대기 오염도, 주관적 토양 오염도의 총 6개 환경관련 문항에 대하여 요인분석을 실시하였다. 요인분석 시, 고유값 1을 기준으로 요인을 추출하였고, 배리맥스(varimax) 방법을 사용하여 요인 회전을 실시하였으며, 회귀분석 방법을 이용하여 요인점수를 추출하였다. 요인분석의 결과 추출된 잠재변수는 표 4.1과 같다.

표 4.1에서와 같이 환경관련 6개 문항에 대하여 요인분석을 실시한 결과 2개의 요인으로

표 4.2: 모형의 오분류율 비교(주관적 수질오염도)

결과변수 \ 정확도		모형 예측 오분류율	모형평가 예측 오분류율
원 변수	주관적 상수도 오염도	0.2592	0.2645
	주관적 하수도 오염도	0.2589	0.2654
잠재 변수	주관적 수질오염도	0.2593	0.2611

표 4.3: 모형의 오분류율 비교(주관적 공간오염도)

결과변수 \ 정확도		모형 예측 오분류율	모형평가 예측 오분류율
원 변수	주관적 소음진동 오염도	0.3784	0.3802
	주관적 악취 오염도	0.3284	0.3186
	주관적 대기 오염도	0.3226	0.3126
	주관적 토양 오염도	0.2512	0.2529
잠재 변수	주관적 공간오염도	0.2917	0.2860

축소되었으며, 1요인의 요인명은 주관적 수질 오염도로 지정하고 2요인의 요인명은 주관적 공간 오염도로 지정하였다.

신경망분석 시, 결과변수를 요인분석에서 추출된 잠재변수인 주관적 수질 오염도와 주관적 공간 오염도로 지정하고, 입력변수를 인구통계학적속성 관련문항으로 지정하였다. 훈련 자료와 모형평가 자료를 각각 1/2로 지정한 뒤, 모형 선택 기준으로 오분류율(misclassfication rate)을 지정하고 신경망의 구조로는 다계층 퍼셉트론으로 지정하여 모형을 생성하였다.

데이터 퓨전 기반 잠재변수에 의한 신경망분석 시, 생성된 모형의 정확도가 기존의 결과변수들에 대한 모형 정확도보다 현저히 떨어진다면 잠재변수에 의한 신경망분석의 모형은 의미가 없을 것이다. 이에 기존의 결과변수에 대한 모형 정확도와 잠재변수에 의한 모형 오분류율을 비교하였다. 기존의 결과변수에 대한 모형 정확도와 첫 번째 잠재변수인 주관적 수질오염도에 대한 모형 오분류율 비교는 표 4.2와 같다. 잠재 변수에 의한 모형의 오분류율은 원 변수에 의한 모형의 오분류율과 큰 차이를 보이고 있지 않다. 특히, 모형평가 예측 오분류율의 경우 원 변수보다 오분류율이 낮아 잠재변수인 주관적 수질오염도에 의한 신경망분석이 효율적임을 알 수 있다.

기존의 결과변수에 대한 모형 정확도와 두 번째 잠재변수인 주관적 공간오염도에 대한 모형 오분류율 비교는 표 4.3과 같다. 잠재 변수에 의한 모형 예측 오분류율과 모형평가 예측 오분류율을 비교한 결과, 원 변수인 지역의 토양 환경오염도의 오분류율 보다는 다소 높게 나타났으나, 지역의 소음진동 환경오염도, 지역의 악취 환경오염도, 지역의 대기 환경오염도에 비해서는 오분류율이 낮아 잠재변수인 주관적 공간오염도에 의한 신경망분석이 효율적임을 알 수 있다.

표 4.4는 첫 번째 잠재변수인 주관적 수질 오염도에 대한 신경망 모형의 결과이다. 주관적 수질 오염도에 대하여 전체적으로 '나쁨'은 39%, '좋음'은 61%로 응답하고 있다. 표 4.4를 구

표 4.4: 주관적 수질 오염도에 대한 근사모형 결과

규칙	노드	결과변수
1	⊙ 학력(대재이상) → 주관적 사회계층(중상류층 이하) → 거주 지역(주거 지역)	나쁨: 77% 좋음: 23%
2	⊙ 학력(고졸이하) → 연령(평균 미만) → 거주 지역(주거 외의 거주 지역)	나쁨: 4% 좋음: 96%
3	⊙ 학력(대재이상) → 주관적 사회계층(중상류층 이하) → 거주 지역(주거 외의 거주 지역)	나쁨: 21% 좋음: 79%

표 4.5: 주관적 공간 오염도에 대한 근사모형 결과

규칙	노드	목표변수
1	⊙ 거주 지역(주거, 상가, 공업지역) → 학력(대재 이하) → 연령(평균 미만)	나쁨: 92% 좋음: 8%
2	⊙ 거주 지역(주거, 상가, 공업지역) → 학력(대재 이하) → 연령(평균 이상)	나쁨: 17% 좋음: 83%
3	⊙ 거주 지역(농, 어촌 지역) → 연령(평균 이하) → 성별(여성)	나쁨: 13% 좋음: 87%

체적으로 살펴보면, 규칙 1에서는 학력이 '대재 이상'이고, 주관적 사회계층이 "중상류층 이하"이면서, 거주 지역이 '주거지역'인 응답자들은 주관적 수질 오염도에 대한 '나쁨'의 응답이 50%로서 전체에 비하여 '나쁨'의 비율이 11% 증가한 것을 알 수 있다. 반면에 규칙 3에서는 학력과 주관적 사회계층은 동일하나 조사 지역이 '주거 외의 거주 지역'인 응답자들은 주관적 수질 오염도에 대한 '나쁨'의 응답이 21%로서 전체에 비하여 18% 감소한 것을 알 수 있어 두 집단에서는 확연한 차이를 보이고 있다.

표 4.5는 두 번째 잠재변수인 주관적 공간 오염도에 대한 의사결정나무 모형 결과이다. 주관적 공간 오염도에 대하여 전체적으로 '나쁨'은 43%, '좋음'은 57%로 응답하고 있다. 표 4.5를 구체적으로 살펴보면, 규칙 1에서는 거주 지역이 '상가지역', '주거지역', '공업지역'이고, 학력이 '대재 이하'이면서, 연령이 '평균 미만'인 응답자들은 주관적 공간 오염도에 대한 '나쁨'의 응답이 92%로서 전체에 비하여 '나쁨'의 비율이 49% 증가한 것을 알 수 있다. 반면에 규칙 2에서는 거주지역, 학력은 동일하나 연령이 '평균 이상'인 응답자들은 주관적 공간 오염도에 대한 '나쁨'의 응답이 17%로서 전체에 비하여 26% 감소한 것을 알 수 있어 두 집단에서는 확연한 차이를 보이고 있다.

본 논문에서 제시하는 방법의 자료분석적 효율성을 파악하기 위하여 원래의 자료인 환경 관련 6개 문항에 대한 의사결정나무 분석 결과를 표 4.6 및 4.7에 제시하였다.

표 4.6은 주관적 수질 오염도의 원래의 변수인 주관적 상수도 오염도, 주관적 하수도 오염

표 4.6: 주관적 상수도, 하수도 오염도의 신경망 모형 결과

규칙	노드	예측	주관적 상수도 오염도	주관적 하수도 오염도
1	⊙ 학력(대재이상) → 주관적 사회계층(중상류층 이하) → 거주 지역(주거 지역)	나쁨	O	O
2	⊙ 학력(고졸이하) → 연령(평균 미만) → 거주 지역(주거 외의 거주 지역)	좋음	O	O
3	⊙ 학력(대재이상) → 주관적 사회계층(중상류층 이하) → 거주 지역(주거 외의 거주 지역)	좋음	O	O

표 4.7: 주관적 소음진동, 악취, 대기, 토양 오염도의 신경망 모형 결과

규칙	노드	예측	주관적 소음 진동오염도	주관적 악취 오염도	주관적 대기 오염도	주관적 토양 오염도
1	⊙ 거주 지역(주거, 상가, 공업지역) → 학력(대재 이하) → 연령(평균 미만)	나쁨	O	O	O	O
2	⊙ 거주 지역(주거, 상가, 공업지역) → 학력(대재 이하) → 연령(평균 이상)	좋음	O	O	O	O
3	⊙ 거주 지역(농, 어촌 지역) → 연령(평균 이하) → 성별(여성)	좋음	O	O	O	O

도의 신경망 모형 결과이고, 표 4.7은 주관적 공간 오염도의 원래의 변수인 주관적 소음진동 오염도, 주관적 악취 오염도, 주관적 대기 오염도, 주관적 토양 오염도의 신경망 모형 결과이다. 표에서 O부분은 원래의 변수에 대하여 규칙이 발견된 것을 의미한다.

표 4.6 및 4.7을 살펴보면, 원래의 자료인 환경관련 6개 문항에 대한 신경망 모형의 규칙(O 표시된 부분)은 총 18개가 나타나고 있다. 주관적 수질 오염도와 주관적 공간 오염도에 대한 신경망 모형의 결과가 주관적 수질 오염도와 주관적 공간 오염도에 대한 원래 문항의 신경망 모형의 결과에 모두 동일하게 나타남을 알 수 있다. 이에 원래의 자료인 환경관련 6개 문항에 대한 신경망 분석을 실시할 경우 총 6개의 신경망 모형을 생성해야 하며, 18개의 규칙을 해석해야하나, 본 논문에서 제시하는 하이브리드 데이터마이닝 방법을 사용하면 2개의 모형만 생성하면 되고 규칙 또한 6개만 해석하면 되므로 신경망 모형 생성 및 모형 해석에 대한 시간을 단축시킬 수 있어 효율적 분석이 가능함을 알 수 있다.

또한 본 논문에서 제시하는 방법과 다른 하이브리드 데이터 마이닝 방법과의 비교 분석을 위하여 표 4.8과 같이 잠재변수에 의한 CART를 이용한 하이브리드 데이터 마이닝의 결과와 잠재변수에 의한 C5.0을 이용한 하이브리드 데이터 마이닝의 결과와의 오분류율을 비교하였다.

표 4.8: 모형의 오분류율 비교(주관적 수질오염도)

하이브리드 데이터마이닝 방법 \ 정확도		모형 예측 오분류율	모형평가 예측 오분류율
잠재변수에 의한 신경망 분석	주관적 수질 오염도	0.2593	0.2611
	주관적 공간 오염도	0.2917	0.2860
잠재변수에 의한 CART 분석	주관적 수질 오염도	0.2731	0.2794
	주관적 공간 오염도	0.3120	0.3004
잠재변수에 의한 C5.0 분석	주관적 수질 오염도	0.2951	0.2847
	주관적 공간 오염도	0.3213	0.3198

표 4.8을 살펴보면, 본 논문에서 제시하는 방법에 의한 주관적 수질 오염도와 주관적 공간 오염도의 모형 예측 및 모형평가 예측 오분류율이 잠재변수에 의한 CART 분석과 잠재변수에 의한 C5.0 분석에 의한 모형의 오분류율 보다 낮음을 알 수 있으므로 본 논문에서 제시하는 하이브리드 데이터 마이닝 방법이 다른 방법들에 비하여 효율적임을 알 수 있다.

5. 결론

현재 통계 조사는 연구의 목적에 따라 조사 문항을 다르게 하여 실시하고 있어 동일한 모집단에 대한 조사라 할지라도 조사의 문항이 다른 경우 각각의 개별적인 분석만 가능한 실정이다. 이는 다른 조사 자료를 활용한다든지 이전 조사 자료와 연계한 분석이 가능하지 못하도록 하는 원인이 되어, 결국에는 고부가가치의 정보를 얻는데 어려움을 주게 된다. 특히 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. 그러므로 본 논문에서 제시한 데이터 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 잠재변수를 이용한 신경망 분석을 적용하는 방안에 대하여 연구하였다. 일반적으로 여러 개의 결과변수를 대상으로 신경망 분석을 실시할 경우, 각각의 모형을 생성하고 이를 분석해야 하므로 많은 시간과 노력이 필요하다. 여러 개의 결과변수에 대한 유사성을 파악하여 몇 개의 변수로 축약할 수 있다면, 신경망의 모형 생성 및 모형 해석에 대한 시간을 단축할 수 있어 효율적인 신경망분석이 가능하며, 이는 적용 사례를 통하여 확인할 수 있었다. 본 논문에서는 퓨전된 데이터에 대한 정확성에 대한 문제와 여러 가지 하이브리드 데이터 마이닝의 결과 비교를 통한 제안된 방법의 효율성 비교에 대한 한계점이 있으며, 향후 퓨전된 데이터의 정확성 문제와 다양한 하이브리드 데이터 마이닝의 적용 결과 비교의 추가적인 연구가 필요하다.

참고문헌

강문식, 이상용 (2002). 데이터 마이닝을 위한 경쟁학습모델과 BP알고리즘을 결합한 하이브리드형 신경망, *Journal of information technology application & management*, **9**, 1–16.

강현철, 한상태, 최종후, 김은석, 김미경 (1999). <Enterprise Miner를 이용한 데이터마이닝>, 자유아카데미.

김만선, 이상용 (2003). 대용량 데이터 처리를 위한 하이브리드형 클러스터링 기법, <정보처리학회논문지>, **10**, 33–40.

김진성 (2003). 연관규칙과 퍼지 인공신경망에 기반한 하이브리드 데이터 마이닝 매커니즘에 관한 연구, <한국경영학회/대한산업공학회 2003 춘계공동학술대회>, 884–888.

박희창, 조광현 (2006). 통계적 데이터 퓨전을 위한 SAS 매크로, *Journal of the Korean Data Analysis Society*, **8**, 1927–1937.

윤경배, 최준혁, 왕창종 (2002). 하이브리드 SOM을 이용한 효율적인 지식 베이스 관리, <정보처리학회논문지>, **9**, 635–642.

조광현, 박희창 (2007). 데이터 퓨전 기반 집단세분화에 의한 의사결정나무 적용 방안, *Journal of the Korean Data Analysis Society*, **9**, 2273–2284.

한상훈, 안일호, 하덕주, 최종후 (2004). 데이터 퓨전과 평가, <한국데이터 마이닝학회 2004 추계학술대회>, 238–254.

황인수 (2002). 데이터 마이닝에서 그룹 세분화를 위한 2단계 계층적 클러스터링 알고리듬, <한국경영과학회지>, **19**, 189–196.

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

[2008년 1월 접수, 2008년 5월 채택]

Application Scheme of Hybrid Data Mining for Fused Data in Statistical Survey

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

Abstract

Today, the statistical survey has been carried out variously for the decision-making and administration of the organization. We use the different items in the statistical survey according to the purpose of study. Currently, Gyeongnam province is executing the social index survey to the provincials every year. But, this survey has the limit of the analysis as execution of the different survey per 3 year cycles. The solution for this problem is data fusion technique. Data fusion is generally defined as the use of techniques that collect to combine data including multiple sources in order to raise the quality of information. But, data fusion doesn't mean the ultimate result. Therefore, efficient analysis for the fused data is also important.

In this study, we suggest the application methodology of neural network by latent variable through the fused data in statistical survey.

Keywords: Data fusion, hybrid data mining, latent variable, neural network.

1) Corresponding author. Professor, Dept. of Statistics, Changwon National University, Changwon, Gyeongnam 641-773, Korea.
E-mail: hcpark@changwon.ac.kr
2) A part-time lecturer, Dept. of Statistics, Changwon National University, Changwon, Gyeongnam 641-773, Korea.
E-mail: cho1023@changwon.ac.kr