데이터 퓨전 : 개념, 문제, 대안

한상훈[1] · 하덕주[2] · 최종후[3]

요 약

최근 마케팅 현업에서 마이크로 마케팅(Micro Marketing)이 마케팅 기법의 화두로 등장하면서 데이터 퓨전(Data Fusion) 또는 데이터 인리치먼트(Data Enrichment)가 각광받는 영역으로 등장하고 있다. 본 연구에서는 데이터 퓨전의 개념과 그를 둘러싸고 있는 통계적 문제와 그 대안에 대하여 논의한다.

주요용어 : 데이터 퓨전(Data Fusion), 대체(Imputation), 평가(Evaluation), 예측력(Predictability), 대표성(Representation)

1. 서론

우리가 살고 있는 시대를 소위, 정보화 시대라고 한다. 정보화 시대의 핵심은 유용한 정보의 활용에 있다. 데이터 홍수시대에 데이터는 사용자에 따라 '황금알을 낳는 거위'가 될 수도 있고 쓰레기가 될 수도 있다. 산재한 데이터를 잘 활용할 수 있는 방법이 데이터 퓨전(Data Fusion)이다. 데이터 퓨전은 데이터 인리치먼트(Data Enrichment), 데이터 매칭(Data Matching) 등 여러 용어로 혼용되고 있다.

최근 마케팅 현업에서 마이크로 마케팅(Micro Marketing)이 마케팅 기법의 화두로 등장하면서 데이터 퓨전이 각광받는 영역으로 등장하고 있다. 본 연구에서는 데이터 퓨전의 개념과 그를 둘러싸고 있는 통계적 문제와 그 대안에 대하여 논의한다.

2장에서는 데이터 퓨전의 개념과 용어에 대해서 알아보고, 3장에서는 데이터 퓨전의 문제와 대안을 그리고 마지막으로 4장에서는 약간의 토의를 덧붙인다.

2. 데이터 퓨전의 개념과 용어

2.1 데이터 퓨전의 개념

데이터 퓨전은 같은 모집단에서 나온 서로 다른 표본들을 포함하는 데이터셋을 합치는 기법 또는 처리과정으로 정의할 수 있다.

영국 National Statistics(2003)의 "National Statistics Code of Practice Protocol on Data Matching"에 따르면 자료 결합(Data Matching)의 종류는 크게 정확 결합(Exact Matching), 판단 결합(Judgemental Matching), 확률적 결합(Probability Matching), 통계적 결합(Statistical Matching), 자료 연결(Data Linking)의 5가지로 구분된다. 궁극적으로 자료 결합

[1] 339-700 충남 연기군 조치원읍 고려대학교 정보통계학과 석사과정, ratm14106@korea.ac.kr

[2] 339-700 충남 연기군 조치원읍 고려대학교 정보통계학과 석사과정, darling79@korea.ac.kr

[3] 339-700 충남 연기군 조치원읍 고려대학교 정보통계학과 교수, jhchoi@korea.ac.kr

을 잘 수행하기 위해서는 이들 방법을 혼용하여 자료 결합을 수행한다.

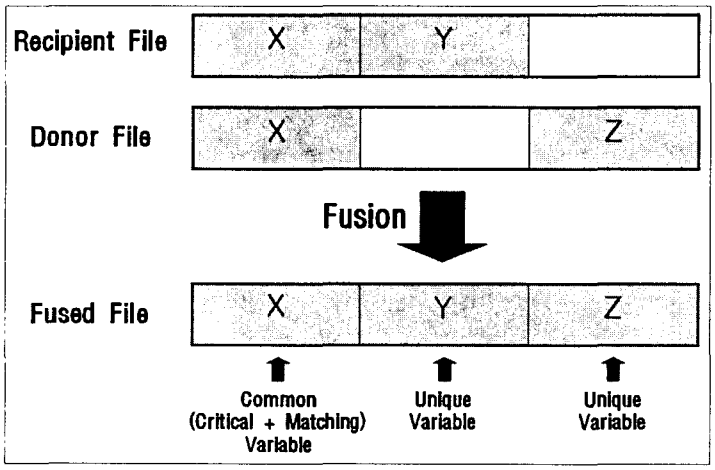
결국 데이터 퓨전의 목적은 계획된 분석의 정보 요구를 만족하게 하기 위해 단일 데이터 소스(Single Data Source)를 풍부하게 하는 것이다. 이러한 의미에서 데이터 퓨전은 데이터 인리치먼트(Data Enrichment)라고도 불리운다[5].

2.2 용어

데이터 퓨전을 설명하기 위해서는 몇 가지 용어들을 정의할 필요가 있다. 그러나 용어의 정의가 완전하게 합의된 상태는 아니다.

두 개의 표본조사 결과를 토대로 데이터 퓨전의 용어를 이해해 보자. <그림 1>에서 보듯이 하나의 조사결과는 2차 조사(Donor Survey)로서 두 번째 조사결과, 즉 1차 조사(Recipient Survey)로 이전될 데이터들을 담고 있다.

이 두 조사에 있어서는 공통되는 여러 질문들이나 기타 정보들이 있어야 하는데, 이들 공통변수(Common Variables)들은 2차 조사에서 어느 응답자들이 1차 조사의 어느 응답자들에게 데이터를 보내는지 결정하는 기초가 된다. 퓨전 절차는 가장 적합한 결합조건을 찾기 위해 2차 조사와 1차 조사 사이의 몇몇 통계학적 유사성 실험(Test of Similarity)에 의존한다. 공통변수들은 유사성을 판단하는 데에 사용되며, 수용 응답자(Recipient)에게 가장 적합한 결합조건이 발견되는 경우, 결합하는 2차 조사로부터 모든 결측 데이터가 이전된다. 따라서 퓨전 기법은 매칭 알고리즘(Matching algorithm)에 따라 설명된다.



<그림 1> Data Fusion Scheme[1]

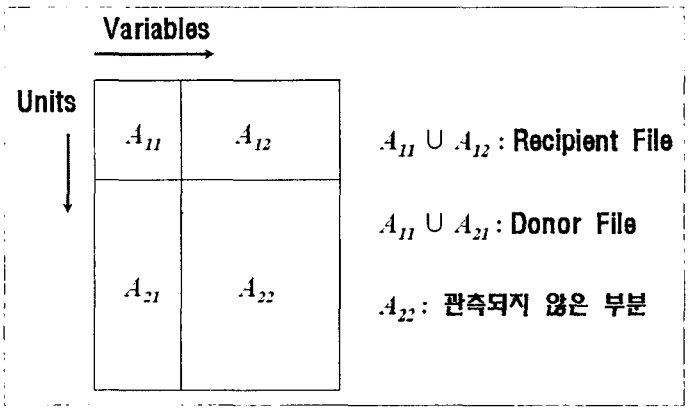
공통변수들은 보통 두 가지 종류로 나누어진다. 우선 기준변수(Critical Variable)라고 함은 성별, 출신지역, 사회적 지위, 혹은 나이 등의 변수를 말한다. 이러한 변수들에 정확한 결합규칙을 부과하는 것은 사실상 표본설계에서의 층화(Stratification)와 매우 유사하다. 만일 기준변수들이 제대로 선정된다면 퓨전의 유효성(Efficiency)은 향상될 것이다. 그 밖의 공통변수들을 결합변수(Matching Variable)라고 부르는데, 기준변수들의 연동에 의해 정의되는 하위그룹 내에서 2차 조사와 1차 조사의 가장 적합한 결합을 찾는 데에 사용된다[4].

3. 데이터 퓨전의 문제와 대안

3.1 데이터 퓨전의 문제

3.1.1 대체(Imputation)

<그림 2>에서와 같이 서로 다른 두 가지 형태의 데이터를 결합하는 경우 A_{22}와 같이 관측되어지지 않는 부분이 발생한다. 퓨전 시 이러한 관측되어지지 않은 부분에 대하여 관심을 가져야하며 이를 위하여 관측하지 못한 A_{22}의 데이터를 수집하거나 추정하여야한다. 이를 추정하는 방법으로 A_{11}과 A_{21}을 이용하여 적절한 모형을 구축하고 A_{12}을 이에 적용하여 A_{22}를 추정하는 대체(Imputation)를 고려함으로써 계획 하에 이루어진 것과 같은 완전한 형태의 데이터를 얻을 수 있다[2].



<그림 2> Recipient File과 Donor File의 결합[2]

3.1.2 평가(Evaluation)

평가라고 함은 퓨전 결과의 유효성의 판단이다. 퓨전의 성과를 측정하는 명백한 통계학적 테스트는 없으며 통계학적인 테스트가 유효성 판단에 대한 접근법으로 필수 불가결한 것도 아니다[4].

3.2 대안

3.2.1 대체 방법(Imputation Method)

결측 데이터(Missing Data)가 포함된 자료의 경우, 다음과 같은 방법들을 이용하여 통계적 분석을 수행하는 것이 일반적으로 알려져 있다. 그러나 이러한 방법들이 결코 상호 배타적인 방법이 아니라 오히려 방법들 사이에는 높은 연관성을 가지고 있다[3].

(1) 결측치를 제거하는 분석방법(Procedure Based on Completely Recorded Units)

분석대상이 되는 통계자료의 일부분이 관측되지 않은 경우 결측치가 포함된 일련의 자료를

분석대상에서 제외한 나머지 완전 자료(Complete Data Set)만을 이용하여 일반적인 통계적 방법을 적용하는 것이다. 방법의 적용이 간단하며, 결측된 부분이 전체에 비하여 비교적 적은 부분이라면, 그 결과 역시 충분히 믿을만하다고 알려져 있다. 그러나 일반적으로 추정된 통계량에서 심각한 편향(Bias)이 있을 가능성이 높으며, 효율성(Efficiency) 역시 크게 떨어진다.

(2) 결측치를 대체하는 분석방법(Imputation-Based Procedure)

분석대상이 되는 통계자료의 결측된 부분을 일련의 적절한 값으로 대체한 후 제공되는 완전 자료를 대상으로 일반적인 통계적 분석방법을 적용하는 것을 일컫는다. 어떠한 방법을 이용하여 결측치를 대체할 것인가에 따라, 평균을 이용하는 방법, 회귀적합을 이용하는 방법 등 여러 가지가 있으며, 다양한 실험 형태에 광범위하게 적용되는 방법이다. 결측치를 하나의 값으로 대체(Single Imputation)하는 방법 외에, 분석대상이 되는 통계자료의 일부분이 관측되지 않은 경우 결측치를 다수의 값으로 대체하는 다중 대체(Multiple Imputation) 방법 등이 있다.

(3) 특정 확률모형에 기반한 분석방법(Model-Based Procedure)

분석대상이 되는 통계자료의 일부분이 관측되지 않은 경우 결측치에 대한 일정한 확률모형을 가정하여 이를 관측된 자료에 대한 확률 분포 모형과의 관련성을 통하여 분석하는 방법이다. 가장 일반적으로 사용되는 대표적인 방법으로 최대우도추정(Maximum Likelihood Estimation) 방법을 이용한 EM 알고리즘(EM Algorithm)등이 있다. 이러한 접근방법은 다양한 형태의 자료에 대하여 유연하게 적용할 수 있다는 장점이 있다.

3.2.2 평가 방법(Evaluation Method)

퓨전의 성과를 평가하는 방법은 예측력과 대표성의 문제로 압축된다[5].

(1) 예측력(Predictability)

기대되는(또는 알고 있는) 목표와 퓨전 결과 사이의 거리(distance) 측도로 예측력을 판단한다.

(2) 대표성(Representation)

퓨전 결과가 원본 수용 파일의 성질을 유지하는가의 문제를 말한다.

4. 토의 및 결론

데이터 퓨전은 이 자체가 하나의 분석이나 결과라기보다는 추후 통계분석결과의 향상을 도모하기 위한 선행작업이라고 할 수 있다. 즉 데이터 퓨전을 통해서 얻은 파일에 의해 추가된 정보를 이용함으로써 분석력을 향상시킬 수 있다.

참고문헌

[1] 안일호 (2003), 혼합형 데이터의 통계적 결합에 관한 연구, 석사학위논문, 고려대학교 대학원.

[2] 안홍덕 (2000), 공식자료의 결합을 통한 기업데이터의 질적 향상, 석사학위논문, 고려대학교 대학원.

[3] 이동희 (1998), 다중대체 방법을 이용한 불완전자료에 대한 판별분석, 석사학위논문, 고려대학교 대학원.

[4] Baker, K. Harris, P. and O'Brien, J. (1989), Data Fusion : An Appraisal and Experimental Evaluation. Journal of the Market Research Society 31 (2) pp. 152~212.

[5] Van Pelt, X. (2001), The Fusion Factory : A Constrained Data Fusion Approach. MSc. Thesis, Leiden Institute of Advanced Computer Science.

[6] National Statistics (2003), National Statistics Code of Practice : Protocol on Data Matching.
www.statistics.gov.uk/about/consultations/general_consultations/downloads/Protocol_on_Data_Matching.pdf