經營科學
第19卷 第1號
2002年 5月

자료융합방법의 성과에 대체수준이 미치는 영향에 관한 연구 : 몬테카를로 시뮬레이션 접근방법*

김성호** · 조성빈*** · 백승익**

Exploring the Effect of Replacement Levels on Data Fusion Methods : A Monte Carlo Simulation Approach*

Sungho Kim** · Sungbin Cho*** · Seung Ik Baek**

■ Abstract ■

Data fusion is a technique used for creating an integrated database by combining two or more databases that include a different set of variables or attributes. This paper attempts to apply data fusion technique to customer relationships management (CRM), in that we can not only plan a database structure but also collect and manage customer data in a more efficient way. In particular, our study is useful when no single database is complete, i.e., each and every subject in the pre-integrated database contains somewhat missing observations. According to the way of treating the common variables, donors can be differently selected for the substitution of the missing attributes of recipients. One way is to find the donor that has the highest correlation coefficient with the recipient by treating common variables metrically. The other is based on the closest distance by the correspondence analysis in case of treating common variables nominally. The predictability of data fusion for CRM can be evaluated by measuring the correlation of the original database and the substituted one. A Monte Carlo Simulation analysis is used to examine the stability of the two substitution methods in building an integrated database.

Keyword : 자료융합, 누락치 추정 및 대체, 몬테카를로 시뮬레이션, 고객관계관리, 데이터관리

논문접수일 : 2001년 7월 30일 논문게재확정일 : 2002년 3월 2일
* 이 논문은 2001년 재단법인 영도 육영회 학술연구지원에 의하여 연구되었음.
** 한양대학교 경영학과
*** 건국대학교 산업공학과

1. 서 론

오늘날 많은 기업들은 새로운 경영전략기법으로서 고객관계관리(Customer Relationship Management : CRM)를 서둘러 도입하고 있다. 고객이 원하는 것이 무엇인지를 파악하고, 그 요구에 맞는 서비스와 제품을 적시에 제공하는 것만이 오늘날 급변하는 시장에서 기업이 생존할 수 있는 유일한 방법일 것이다. 그래서, 많은 기업들은 고객의 욕구를 분석하고, 인터넷 상에서 다양한 마케팅 전략을 수행할 수 있는 정보시스템 도입에 막대한 노력과 투자를 하고 있는 실정이다. 그럼에도 불구하고 CRM을 위한 이러한 기업의 노력과 투자는 기업의 영업활동 성과에는 그다지 크게 영향을 미치지 못하고 있다. 효율적인 CRM을 위해서는 고객과 관련된 모든 정보를 획득하고, 그것을 관리하고 활용할 수 있는 기업의 능력이 필수 요건일 것이다. 그러나, 고객에 대한 전체적이고도 명확한 그림을 그리기 위해 필요한 고객정보를 획득하고 관리하는 과정에는 상당한 비용과 시간이 소요된다. CRM을 성공적으로 추진하기 위해서는 고객정보의 양보다는 질적인 문제가 더욱 중요시된다. 만일 데이터가 고객에 대한 최신의 정보를 정확하게 반영하지 못한다면 이를 기반으로 한 의사결정은 오히려 역효과를 가져올 수도 있다. CRM을 서둘러 도입하고 있는 많은 기업들은 고객분석의 입력 자료가 되는 고객정보를 획득하고 관리하는 과정에 대해서는 상대적으로 적은 관심을 보이는 경향이 있다. 고객 데이터를 분석하고, 서비스를 제공할 수 있는 정보시스템 도입에 앞서 기업 내 · 외에 산재해 있는 고객정보의 수집과 융합이 먼저 수행되어져야 할 것이다.

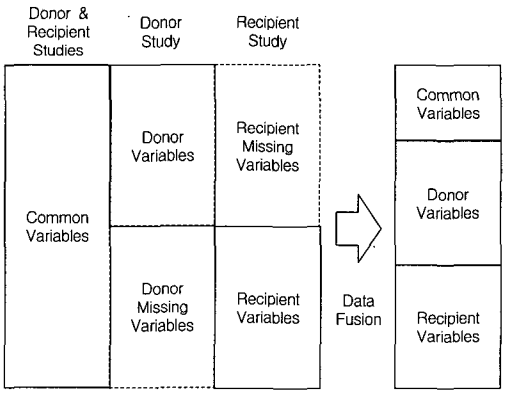
많은 기업들은 다양한 채널을 통하여 고객에 대한 정보를 지속적으로 수집하고, 그것을 분석하여 고객 개개인을 위한 마케팅 전략을 수립하고 있다. 예를 들어, 고객의 과거 구매자료, 인구 통계학적인 자료, 인터넷을 방문하였다면 어떤 사이트를 자주 방문하였는지에 대한 자료(로그파일)를 분석하여 고객 개개인의 욕구를 파악하고 있다. 고객정보를 수집하기 위한 여러 가지 방법 중에서 가장 보편적이고, 쉽게 사용할 수 있는 방법은 설문지일 것이다. 그러나, 설문지를 통하여 획득하는 정보의 질과 설문지 설계, 배포, 그리고 회수를 위하여 투자하는 기업의 비용과 노력을 비교한다면 설문지는 고객정보를 수집하기 위한 효율적인 방법은 결코 아닐 것이다. 특히, 인터넷이 보편화되면서 많은 기업들은 설문지 배포와 회수에 드는 비용과 노력을 줄이기 위해서 인터넷을 사용하여 설문지 조사를 하고 있다. 그러나, 많은 연구에서 인터넷 설문지를 통하여 얻은 정보의 신뢰성 문제를 지적하고 있다. 고객의 욕구를 파악하기 위해서 어떤 과학적인 정보 수집 방법을 이용하더라도 완전한 고객 정보를 수집하기는 현실적으로 불가능하다. 응답자의 불성실한 답변이나 자료 처리과정에서 생기는 오류로 누락치가 생길 수 있을 것이다. SPSS나 SAS와 같은 통계 소프트웨어에서는 누락되어진 항목이 발견되어질 시에 표본 집단 내에 있는 다른 응답자들의 그 항목에 대한 평균값으로 대체하는 간단한 방법을 사용하고 있다. 잘못된 누락치에 대한 추정은 정보의 오류를 발생시킬 수도 있을 것이다. 본 연구에서는 이렇게 자료에 누락치가 발생하였을 시에 그 값을 추정할 수 있는 좀 더 체계적인 방법으로 자료융합(Data Fusion) 방법을 제시하고, 그 방법의 유용성을 실증적으로 평가하는데 그 주요 목적을 두고 있다.

2. 자료융합(Data Fusion)

자료융합(Data Fusion)은 두 개 이상의 표본집단으로부터의 서로 다른 설문지 조사 자료를 융합하여 하나의 융합 되어진 자료를 만들어 내는 과정이다(Kamakura & Wedel, 2000). 이 방법은 커뮤니케이션 분야에서 누락되지 않은 정보를 기초로 누락된 정보(Missing Value)를 추정하는 하나의 방법으로서(Baker et al., 1989), 주로 다차원 척도법(Multidimensional Scaling : Kruskal & Wish,

1978)의 특별한 경우인 Correspondence Analysis를 사용하여 누락치를 추정하는 방법이다. 자료융합은 커뮤니케이션 분야에서 뿐 만 아니라 설문지를 통한 마케팅 조사 등에서 빈번하게 발생하는 누락치를 추정하는 데에도 유용하게 사용될 수 있다(Kamakura & Wedel, 2000).

자료융합에서는 응답자 표본을 기증자(Donor)와 수혜자(Recipient)로 구분한다. 기증자란 어떤 특정한 수혜자에게 그가 가지고 있는 누락치에 대한 대체값(추정값)을 제공해 주는 응답자를 말하며, 수혜자란 기증자로부터 자신이 가지고 있는 누락치에 대한 추정값을 제공받는 응답자를 말한다. 자료융합에서 각각의 수혜자는 우선 자기로부터 가장 가까운 거리에 있는 기증자를 찾는다. 기증자를 찾는 방법은 p 개의 공통변수(Common Variable)로 구성된 p 차원의 공간에서 특정한 수혜자로부터 다른 응답자들(잠재적 기증자)과의 거리를 계산하여 그 수혜자로부터 가장 가까운 거리에 위치한 응답자를 기증자로 선택하게 된다. 일반적으로 공통자료로 응답자의 인구 통계적 정보(Demographic Information)를 사용한다. 일단 특정 수혜자가 자신으로부터 가장 가까운 거리에 있는 기증자를 찾으면 그 기증자가 가지고 있는 변수들의 값이 수혜자의 누락치에 대한 추정치가 된다. 만일 이 기증자 역시 수혜자와 마찬가지로 특정한 변수에 대한 누락치를 지니고 있으면 그 다음으로 가까운 거리에 있는 응답자를 수혜자로 찾는다. 수혜자와 기증자와의 거리를 계산하기 위해서 일반적으로 상관계수(Correlation Coefficient)와 Correspondence Analysis를 사용할 수 있다. 만약 수혜자나 기증자의 비누락치(Non-Missing Value)로 구성된 공통자료가 응답항목별 자료(Categorical Variables)라면 어떤 특정한 차원을 지닌 공간 내에서 각각의 수혜자에 대한 기증자를 파악하기 위하여 응답자간의 거리(예를 들면 Euclidean Distance나 Mahalanobis Distance), 혹은 응답자간의 상관계수를 직접적으로 계산하는 것은 불가능하다. 이러한 경우, 응답자간의 거리를 계산하기 위하여 사용되는 것이 Correspondence Analysis이다. <그림 1>은 자료융합 과정을 설명해 주고 있다. 기존의 연구에서는 응답자의 부주의 혹은 무능력으로 발생한 비의도적인 자료의 누락치(Unintentional Data Missing)를 예측하거나 대체하는 여러 가지 방법을 제안하고, 실증적으로 그 성과를 비교하였다(Downey & King, 1998 ; Kromrey & Hines, 1994 ; Landerman et al., 1997). 그러나, 공통변수를 미리 정해놓은 의도적인 누락치(Intentional Data Missing)의 대체값 혹은 예측값을 제공해 주는 방법들에 대한 연구는 극히 드물다.(Baker et al., 1997, Kamakura & Wedel, 2000). 본 연구에서는 상관계수를 이용하여 기증자를 찾는 자료융합 방법과 Correspondence Analysis를 이용하여 기증자를 찾는 자료융합 방법의 정확도를 비교하는데 그 주요 목적을 두고 있다.



〈그림 1〉 자료융합 과정

2.1 자료융합의 활용

자료융합 방법은 주로 마케팅 분야에서 많이 사용되고 있다. 자료융합 방법은 크게 세 가지 분야에서 널리 활용되어지고 있다.

2.1.1 시장조사(Market Research)

자료융합 방법을 활용하여 시장조사에 필요한 여러 가지 자료를 융합하여 하나의 자료를 만들어 낼 수가 있다. 서로 다른 분야에 대한 조사일 지라

도 공통된 변수나 유사한 변수가 존재한다면 자료융합 방법을 사용하여 직접 조사하지 않더라도 그 표본 집단의 의견을 조사할 수가 있다. 예를 들어, 소비재를 생산하는 기업에서는 고객구매이력에 관한 구매 및 성향분석 데이터베이스를 광고에 관련된 데이터베이스와 결합시킴으로써 그들의 시장에 관한 좀 더 확장된 시각을 가질 수 있을 것이다. 이 과정에서 기업일선의 대부분 관리자들은 인구통계학적 자료를 이용하여 묵시적으로 자료를 융합 시키고 있음을 주목할 필요가 있다. 어떠한 제품이나 서비스에 대한 태도차이를 연령에 의하여 분석하고자 할 경우에, 두 가지 이상의 상이한 자료들을 종합하여 결론을 도출하는데 이는 자료융합의 대표적인 사례이다. 체계적이고 과학적인 방법론을 구축하고 적용함으로써 관리자들이 묵시적인 자료융합에서 발생할 수 있었던 고객분석에 대한 오류를 감소시키고 정확한 분석과 접근을 통하여 고객의 만족도를 향상시킬 수 있을 것이다.

2.1.2 Direct Marketing

둘째로는 자료융합을 사용함으로 해서 여러 가지 종류의 물건에 대한 선호도 조사 자료를 융합함으로 해서, 고객이 하나의 물건에 대하여 직접 설문지에 답하지 않았을 지라도 그 물건에 대한 다른 응답자의 선호도를 기초로 하여 그 고객의 선호 정도를 예측할 수 있다. 자료융합을 통하여 추정된 고객의 선호도를 기초로 하여 선별적으로 고객에게 Direct Marketing을 실시함으로 해서 그것의 효율성을 높일 수 있다. 호주에서는 은행의 고객정보시스템에 고객이 관심 있어 할만한 매력적인 금융상품을 추천하여 교차판매(Cross Selling)를 증가시키는데 자료융합방법을 이용하고 있다.

2.1.3 Media Planning

셋째로는 자료융합 방법을 사용하여 고객에 대한 구매행동에 관한 자료와 TV 프로그램에 대한 시청률 조사 자료를 융합하여 광고나 판촉을 기획할 시에 기본적인 자료로 많이 사용하고 있다. 즉 구매행동 자료와 시청률 자료를 융합함으로 해서 9시 뉴스를 즐겨 보는 사람이 어떤 물건을 자주 구입하는지를 추정할 수가 있을 것이다. 또한 커피를 자주 구입하는 사람은 어떤 TV 프로그램을 선호할 지를 추정할 수도 있을 것이다. 자료융합 방법을 사용하지 않는 다면 구매패턴 조사를 위한 설문지와 TV 프로그램 선호도를 조사하기 위한 두 개의 설문지를 동일한 표본 집단에 배포하여 조사하여야 할 것을 자료 융합을 사용함으로 해서 이와 같이 서로 상이한 성격을 가지고 있는 자료를 융합하여 체계적인 광고나 판촉 기획에 도움을 줄 수가 있다.

2.2 상관계수와 Correspondence Analysis를 이용한 자료융합 방법

자료융합 방법에서는 누락치를 추정하기 위해서 누락된 항목을 포함하고 있는 응답자와 가장 비슷한 속성을 가지고 있는 응답자의 그 항목에 대한 값으로 대체함으로써 누락치를 측정해 낸다. 자료융합에서는 크게 두 가지 방법을 사용하여 가장 가까운 응답자를 찾고 있다.

2.2.1 상관계수를 이용한 자료융합(방법 I)

상관계수를 이용한 자료융합 방법은 응답자들의 공통 변수들의 상관계수를 기초로 하여 누락된 값을 가지고 있는 응답자와 가장 가까운 속성을 가진 응답자를 찾아내고 그 응답자로부터의 값을 가지고서 누락된 값의 추정치로 대체하는 방법이다. 상관계수를 사용하여 응답자 간의 거리를 구하기 위해서는 모든 공통 변수가 계량변수(Metric Variable)라고 가정을 해야 한다. 누락된 변수의 추정치는 누락치가 발견된 응답자 I(수혜자)와 가장 상관계수가 높은 응답자 j(기증자)의 변수 값으로 대체하여 누락치를 추정하였다. 만약 가장 상관계수가 높은 응답자 j도 같은 변수의 값이 누락되어 있으면 그 다음 상관계수가 높은 응답자 k의 그 변수의 값으로 대체한다.

응답자 i와 j의 상관계수(i≠j) :

$$r_{ij} = \frac{\sum_{k=1}^{n} (\overline{Y}_{ik} - Y_{ik})(\overline{Y}_{jk} - Y_{jk})}{\sqrt{\sum_{k=1}^{n} (\overline{Y}_{ik} - Y_{ik})^2} \sqrt{\sum_{k=1}^{n} (\overline{Y}_{jk} - Y_{jk})^2}}$$

where, y_{ik} : 응답자 i의 k번째 변수
i = 1, …, n 응답자
k = 1, …, n : 공통변수

2.2.2 Correspondence Analysis을 이용한 자료융합(방법 II)

Correspondence Analysis(Hoffman and Franke 1986 ; Carroll et al., 1986, 1987)란 다차원 척도법(Multidimensional Scaling ; Carroll & Arabie 1980)의 일종으로서 분석자료의 종류에 있어서 일반 다차원 척도법과 구별되는 분석 기법이다. 일반적으로 다차원 척도법에 사용되는 입력자료의 종류는 다차원 척도법의 유형에 따라 등간척도 혹은 비율척도를 사용한 Rating Data(Metric MDS의 경우)와 서열척도를 사용한 Ranking Data(Non-Metric MDS의 경우)로 나뉘어진다. 이에 비하여 Correspondence Analysis의 경우는 입력자료가 Dummy변수를 포함한 명목 척도이거나 혹은 응답항목(Dichotomous 이거나 Multichotomous) 에 따른 응답의 빈도수(frequency)라는 점에서 일반적인 다차원 척도법과 구분되어진다. 따라서 Correspondence Analysis란 일반적으로 N-Way Contingency Table 혹은 Cross-Tabulation Table을 분석하는데 사용되는 기법이다.

Correspondence Analysis 역시 다른 다차원 척도법과 마찬가지로 입력자료에 나타난 개체(소비자, 제품, 기업, 제품의 사용상황 등)들을 몇 차원의, 일반적으로 p-차원의 공간에 점(point)으로 나타내는 기법이다. 이 과정에서 입력자료에 나타난 개체들간의 상대적인 관계를 Correspondence Analysis를 통하여 구성한 공간에서도 동일하게 유지하여 나타낸다는 것이 Correspondence Analysis(다른 다차원분석기법과 마찬가지로)의 특징이라고 할 수 있다. 또한 Correspondence Analysis에서는 입력자료의 가로와 세로에 나타난 개체들을 동시에 동일한 공간에 점으로 나타내는 Joint Space 분석기법의 일종이라고 할 수 있다. 또한 Correspondence Analysis의 알고리즘에 따라서는 가로와 세로의 개체들간의 거리가 직접 비교가능할 수도 있다(Carroll et al., 1986 ; 1987).

Carroll et al.,(1986, 1987)의 알고리즘에 따르면, 우선 분석의 대상이 되는 Contingency Table혹은 Cross Tabulation Table matrix F행렬(F는 가로 i줄과 세로 j줄로 되어 있다고 하자)를 다음과 같이 H행렬로 정상화(normalize)한다.

$$H = R^{-1/2} F C^{-1/2}$$

R행렬은 i x i diagonal matrix이며 C행렬는 j x j diagonal matrix이다. 이들 R과 C는 각각 가로와 세로의 합계의 제곱근의 역수(Reciprocals of the Square Roots of Row and Column Marginal)로 구성되어 있다. 다음으로 H행렬은 다음의 식을 사용하여 chi-square 거리척도로 전환된다.

$$H = P \Delta Q'$$

여기에서 $P'P = Q'Q = I$이며 Δ는 Diagonal Metric이다. 끝으로 p-차원상에서의 가로줄(X)과 세로줄(Y)에 나타난 개체의 좌표는 각각 다음과 같다.

$$X = R^{-1/2} P(\Delta + I)^{1/2}$$
$$Y = X^{-1/2} R(Q + I)^{1/2}$$

상관계수를 이용한 자료융합과는 달리 Correspondence Analysis를 사용한 자료융합 방법에서는 공통 변수를 범주형 변수라고 가정하여야 한다. 본 연구에서는 범주형 공통 변수에 기초하여 SAS의 Correspondence Analysis를 적용하여 각 응답자간의 거리를 계산하였다.

응답자 i와 j의 거리(i≠j) :

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (d_{ik} - d_{jk})^2}$$

where, d_{ik} : 응답자 i의 k 차원의 좌표.

i = 1, ···, n : 응답자

k = 1, ···, n : 차원

누락된 변수의 값은 누락치가 발견된 응답자 i (수혜자)와 가장 거리가 가까운 응답자 j(기증자)의 변수의 값으로 대체하여 예측한다. 상관계수를 이용한 자료융합 방법과 마찬가지로 만약 가장 거리가 가까운 응답자 j도 같은 변수의 값이 누락되어 있으면 그 다음 거리가 가까운 응답자 k의 변수의 값으로 대체하고, 만약 응답자 k도 변수의 값이 누락되어 있으면, 이 과정은 변수의 값이 누락되지 않고 거리가 다음 순서로 가까운 응답자 l에게까지 확장한다.

2.3 연구목적

본 연구의 구체적인 연구 목적은 다음과 같다.

- 상관계수를 사용한 자료융합과 Correspondence Analysis를 사용한 자료융합을 통하여 얻어진 누락치에 대한 추정치는 얼마나 본래 값에 가까운가(즉, 추정치의 신뢰도)?
- 삭제된 속성의 수가 추정치의 신뢰도(정확도)에 미치는 영향은 어느 정도인가?

3. 연구방법

상관계수를 이용한 자료융합방법과 Correspondence Analysis를 이용한 자료융합방법을 각각 이용하여 기증자로부터 구한 추정치가 삭제하기 전의 수혜자의 값(Original Value)에 얼마나 가까운가를 평가하기 위하여 본 연구에서는 다음의 두 가지 척도를 사용하였다.

3.1 상관계수(Correlation Coefficient)를 이용한 평가

수혜자가 지니고 있던 본래의 값(Original Value)과 자료융합을 통하여 구한 추정치(Recovered Value)간의 상관계수를 각 응답자의 수준에서 구하였다. 두 개의 자료간에 상관계수가 높으면 자료융합방법을 이용하여 추정한 값이 본래의 값에 근접 하다는 것을 의미한다.

3.2 Adjusted Rand Index(ARI)를 이용한 평가

수혜자의 본래의 값으로 구성된 자료와 자료융합을 통해 예측되어진 추정치로 구성되어진 자료를 각각 군집분석을 한 후, 각각의 자료를 사용하여 구한 군집 소속간의 일치도(i.e., 세분시장의 안정도)를 자료융합의 평가척도로 사용하였다. 군집간의 일치도를 구하는 척도로서 많이 사용되는 것은 Rand Index이다. Rand Index란 군집에 속해 있는 응답자 쌍들의 빈도수의 비율이다. Rand Index의 분자는 동일한 두 명의 응답자 쌍이 동일한 군집에 소속되어 있는가 혹은 상이한 군집에 소속되어 있는가의 빈도 수이며 분모는 전체 응답자 쌍의 수이다. 만약 두 개의 군집이 정확하게 일치한다면 Rand Index는 1.0이다. 만일 군집의 구성원들간에 일치가 전혀 이루어지지 않는다면 Rand Index는 0이다. 그러나 본래의 Rand Index에는 상향적 오차

〈표 1〉 Rand Index와 Adjusted Rand Index

True Structure (알려져 있는 군집구조)	Test Structure (군집분석을 통하여 발견한 군집구조)		
	동일군집에 속한 개체의 쌍	상이한 군집에 속한 개체의 쌍	합
동일군집에 속한 개체의 쌍	A	B	A + B
상이한 군집에 속한 개체의 쌍	C	D	C + D
합	A + C	B + D	R

가 존재하므로 그 오차를 수정하여 많이 쓰이고 있는 것이 Adjusted Rand Index(ARI)이다. Rand Index가 높을수록 자료융합을 통해 본래의 값을 정확하게 추정한 것이다. <표 1>는 Rand Index와 ARI를 구하는 식을 보여 주고 있다.

Original Rand Index :

$$\frac{(A+D)}{R} = \frac{(A+D)}{\frac{1}{2}N(N-1)}$$

Adjusted Rand Index

$$\frac{R(A+D)-[(A+B)(A+C)+(C+D)(B+D)]}{R^2-[(A+B)(A+C)+(C+D)(B+D)]}$$

$$R = \frac{1}{2}N(N-1)$$

3.3 평가 자료

본 연구에서 두 가지의 자료융합 방법에 의한 누락치 추정에 대한 정확도를 비교하기 위하여 다음의 설문지 자료를 사용하려고 한다. 본 연구에서 사용한 자료는 600명의 응답자들로부터 수집된 자동차 딜러쉽에 대한 선호도 조사 자료이다. 구체적으로, 본 연구에서 사용된 자료는 13개의 속성으로 구성되어 있으며 각 속성의 수준은 2개에서 8개이다(<표 2> 참조). 두 가지의 자료융합 방법을 이용한 누락치들에 대한 추정치의 정확도를 평가하기 위하여 본 연구에서는 13개의 속성과 그에 속한 총 49개의 속성수준 자료를 이용하였다(<표 2> 참조). 처음의 6가지 속성(속성변수 Y_1에서 Y_6까지에 해당됨)은 600명의 응답자 자료를 자료융합을 위한 공통변수로 사용하고, 나머지 7가지 속성(속성변수 Y_7에서 Y_{13}까지에 해당됨)은 각각 200명으로 구성된 세 개의 집단으로 나누고, 각 집단마다 누락한 속성의 수를 달리하였다. 즉,

- 집단 A : 7가지 속성 중 무작위로 하나의 속성을 선택, 그에 해당되는 속성수준 모두를 누락시킨다.
- 집단 B : 7가지 속성 중 무작위로 두 개의 속성을 선택, 그에 해당되는 속성수준 모두를 누락시킨다.
- 집단 C : 7가지 속성 중 무작위로 세 개의 속성을 선택, 그에 해당되는 속성수준 모두를 누락시킨다.

〈표 2〉 자동차 딜러 쉽의 속성과 속성수준의 수

속 성	속성변수	속성수준의 수	속성수준 변수
판매차종	Y_1	6	Y_{1-1}, Y_{1-2}, Y_{1-3}, Y_{1-4}, Y_{1-5}, Y_{1-6}
전시장(매장)위치	Y_2	3	Y_{2-1}, Y_{2-2}, Y_{2-3}
전시장(매장)내부	Y_3	4	Y_{3-1}, Y_{3-2}, Y_{3-3}, Y_{3-4}
차량구입 도우미	Y_4	3	Y_{4-1}, Y_{4-2}, Y_{4-3}
할부판매	Y_5	3	Y_{5-1}, Y_{5-2}, Y_{5-3}
구매거래	Y_6	3	Y_{6-1}, Y_{6-2}, Y_{6-3}
보상판매	Y_7	4	Y_{7-1}, Y_{7-2}, Y_{7-3}, Y_{7-4}
전문영업사원	Y_8	3	Y_{8-1}, Y_{8-2}, Y_{8-3}
신차서비스	Y_9	4	Y_{9-1}, Y_{9-2}, Y_{9-3}, Y_{9-4}
렌탈카	Y_{10}	2	Y_{10-1}, Y_{10-2}
서비스플랜	Y_{11}	3	Y_{11-1}, Y_{11-2}, Y_{11-3}
부 품	Y_{12}	3	Y_{12-1}, Y_{12-2}, Y_{12-3}
가격할인	Y_{13}	8	Y_{13-1}, Y_{13-2}, Y_{13-3}, Y_{13-4}, Y_{13-5}, Y_{13-6}, Y_{13-7}, Y_{13-8}

세 개의 집단에서 누락된 값을 추정하기 위해서 상관계수와 Correspondence Analysis를 이용한 자료융합 방법을 각각 사용하여 누락된 값을 추정하였다.

Y_7에서 Y_{13}까지의 속성에 해당하는 27개의 속성수준(즉, Y_{7-1}에서 Y_{13-8}까지의 속성수준)의 원래자료와 위의 두 가지 방법에 의하여 누락치를 예측하여 보완함으로써 누락치가 없어진 예측된 자료를 이용하여 두 자료의 일치도를 위에서 언급한 바와 같이 상관계수와 Rand Index를 사용하여 평가하였다. 연구결과의 안정성을 시험하기 위하여 Monte Carlo Simulation을 실시하였다. 본 연구에서는 자동차 딜러쉽에 대한 자료를 사용하여 두 가지 다른 누락치 추정 방법의 정확도를 비교하였다. <표 2>는 평가자료로 사용할 고객자료의 구조를 설명해 주고 있다.

4. 연구결과

Y_7에서 Y_{13}까지의 속성에 해당하는 27개의 속성수준(즉, Y_{7-1}에서 Y_{13-8}까지의 속성수준)의 원래자료와 위의 두 가지 방법에 의하여 누락치를 예측하여 보완함으로써 누락치가 없어진 예측된 자료를 이용하여 두 자료의 일치도를 다음과 같이 측정하였다.

4.1 상관계수에 의한 평가 결과

각 응답자별로 두 자료간의 상관계수를 구한다. 즉, 각 집단별로 응답자 200명에 대한 200개의 상관계수를 구한 후 이에 대한 평균상관계수를 계산한다.

집단의 평균상관계수 :

$$\overline{R} = \frac{\sum_{i=1}^{200} R_i}{200}$$

where, R_i : 응답자 i의 원래자료와 예측된 자료와의 상관계수

$i = 1, 2, \cdots, 200$: 각 집단내의 응답자

연구결과의 안정성을 시험하기 위하여 Monte Carlo Simulation을 실시한다. 100회의 시뮬레이션을 실시한 결과 평균상관계수의 평균은 <그림 2>에서 보는 봐와 같이 일정한 값에 수렴하고 있음을 알 수 있다.

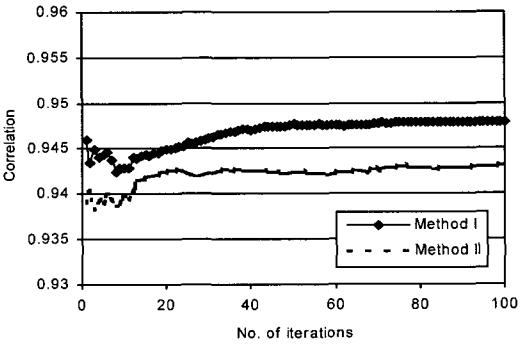
평균상관계수의 평균 :

$$\overline{\overline{R}} = \frac{\sum_{i=1}^{100} \overline{R}^{(i)}}{100}$$

where.

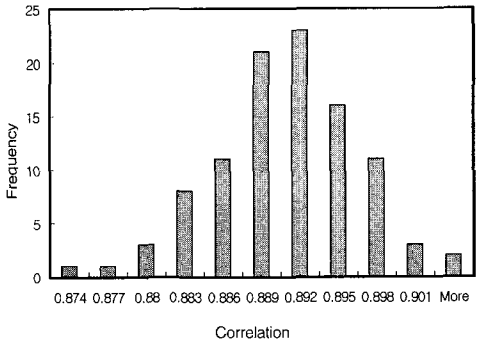
$\overline{R}^{(i)}$: i번째 시뮬레이션의 평균상관계수

$i = 1, 2, \cdots, 100$: 시뮬레이션 횟수



(집단 A 하나의 속성을 누락한 집단-에 대하여 방법 I과 방법 II를 적용한 결과)

〈그림 2〉 평균상관계수의 수렴여부의 예



(집단 B에 대하여 방법 I을 적용한 결과)

〈그림 3〉 시뮬레이션결과에 따른 평균상관계수의 히스토그램의 예

또한 각 집단에 대한 평균상관계수의 분포도 <그림 3>에서 보는 봐와 같이 대략 정규분포의 모양을 따르고 있어 평균상관계수의 평균값을 대표 값으로 사용할 수 있었다.

<표 3>는 Monte Carlo Simulation에 의한 평균상관계수의 분포특성을 요약하고 있다. 누락된 속성이 하나인 경우 평균상관계수의 평균은 두 방법 모두 0.94를 상회하여 매우 높았다. 누락된 속성이 두 개인 집단 B에 대한 예측은 상관계수가 0.89에 근접하여 역시 매우 정확한 예측이 이루어지고 있음을 보여주고 있다. 마지막으로 누락된 속성이 3

〈표 3〉 Monte Carlo Simulation에 의한 평균상관계수의 분포특성

		평 균	표준편차
집단 A	(삭제된 속성의 수 1개)	0.9455	0.0049
집단 B	(삭제된 속성의 수 2개)	0.8904	0.0057
집단 C	(삭제된 속성의 수 3개)	0.8453	0.0073
방법 I		0.8933	0.0436
방법 II		0.8941	0.0392
집단 A	방법 I	0.9479	0.0040
	방법 II	0.9431	0.0045
집단 B	방법 I	0.8899	0.0055
	방법 II	0.8909	0.0058
집단 C	방법 I	0.8422	0.0060
	방법 II	0.8484	0.0072

〈표 4〉 Monte Carlo Simulation에 대한 상관계수의 분산분석 결과

Source	d.f.	Sum of square	Mean square	F value	Pr > F
Model	5	1.0104	0.2021	6432.04	< 0.0001
집단	2	1.0072	0.5036	16029.2	< 0.0001
방법	1	0.0001	0.0001	3.18	0.0749
집단*방법	2	0.0031	0.0015	49.26	< 0.0001
Error	594	0.0187	< 0.0001	-	-
Total	599	1.0290	-	-	-

개인 경우에도(즉, 누락율 약 43%인 경우) 평균상관계수가 0.84 이상으로써 두 방법에 의한 예측이 상당히 정확도가 높음을 보여준다. 예상했던 바와 같이 삭제된 속성의 수가 증가할수록 본래의 값과 추정 값간의 상관계수는 낮아진다는 것을 보여주고 있다. 또한 표준편차는 모두 매우 낮아 예측이 상당히 안정적임을 보여주고 있다. 누락된 속성의 수가 증가함에 따라 표준편차가 점차 증가함을 확인할 수 있다. <표 4>에서는 누락된 속성의 수는 평균상관계수에 유의적 차이를 가져오고 있으며, 두 가지 융합방법 자체는 비유의적 영향을 미치고 있으며, 누락된 속성의 수와 두 가지 다른 융합방법은 교호효과(interaction effect)가 존재하고 있음을 보여주고 있다. 누락된 속성의 수가 1개일 때는 상관계수에 의한 방법이 좋은 예측을 보여주며, 누락된 속성의 수가 증가하여 두 개 이상일 경우에는 Correspondence Analysis를 사용하여 거리를 구하고 예측한 방법 II가 더 좋은 예측을 보여주고 있다. 방법 II가 방법 I에 비하여 평균상관계수의 분포에 대한 편차가 큼을 또한 주목할 수 있을 것이다.

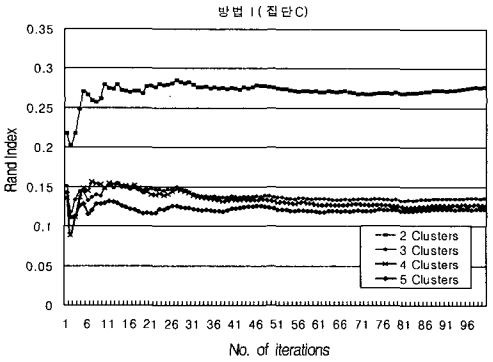
4.2 Adjusted Rand Index에 의한 평가 결과

본 연구에서는 자료융합의 신뢰도를 측정하기 위한 또 하나의 평가기준으로서 세분시장의 안정도를 사용하였다. 세분시장의 안정도를 실험하기 위해서 본 연구에서는 Adjusted Rand Index을 사용하여 누락된 값을 추정하여 새로이 구성한 자료가 얼마나 정확하게 본래의 자료의 군집 구조를 도출하였는지를 살펴 보았다. 자료융합 방법이 궁극적으로 추구하는 목적은 누락된 각각의 값을 정확하게 추정하는 것일 것이다. 그러나, 현실적으로 누락된 값을 100% 정확하게 예측하는 것은 거의 불가능하다. 그래서, 본 연구에서는 두 개의 자료융합 방법이 얼마나 정확하게 본래의 자료의 군집 구조를 도출하였는지를 기준으로 하여 자료융합 방법의 성과를 비교함으로써 좀 더 현실적인 평가척도에 의해서 각 방법을 평가하였다.

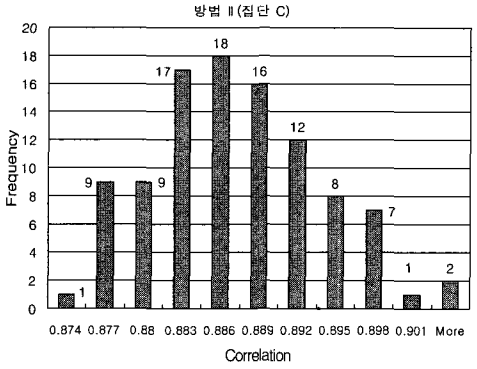
본 연구에서는 현재 가장 많이 군집분석 방법으로 활용되고 있는 K-평균 군집분석 프로그램 중 하나인 Convergent Cluster Analysis(CCA) 프로그램을 사용하여 추정된 자료와 본래의 자료의 군집구조를 도출하였다. 두 자료의 군집구조를 보다 정확하게 비교하기 위하여 CCA를 사용하여 2개에서부터 5개 사이의 군집을 각각 조성하고, 각 군집구조의 일치도를 측정하여 자료융합 방법에 의한 누락치 추정의 정확도를 평가하였다.

<그림 4>에서는 집단 C(네 개의 속성을 누락시킨 집단)에 대하여 방법 I(상관계수를 이용한 자료융합)을 사용하여 누락된 속성을 추정하였을 때 얼마나 정확하게 추정하였는지를 Monte Carlo Simulation을 사용하여 평가한 그림이다. 상관계수를 사용한 자료융합의 시뮬레이션 결과에서와 같이 일정한 값에 Adjusted Rand Index가 수렴함을 알 수 있다. 또한 각 집단에 대한 Adjusted Rand Index의 분포도 <그림 5>에서 보는 바와 같이 대략 정규분포의 모양을 따르고 있어 평균 RAND Index의 평균값을 대표 값으로 사용할 수 있었다. 이와 같은 결과는 집단 C에서 뿐만 아니라 집단 A와 집단 B의 실험에서도 동일한 결과를 얻을 수 있었다.

<표 5>는 Monte Carlo Simulation에 의한 평균 Adjusted Rand Index의 분포특성을 요약하고 있다. 상관계수를 이용한 평가에서 관측한 바와 같이 삭제된 속성의 수가 증가할수록 Adjusted Rand Index는 현저하게 저하되고 있음을 관측할 수 있다. 그리고, 표준편차는 상관계수에 의한 평가 결과와 비교하여 상대적으로 높으나, 상당히 안정적



〈그림 4〉 Adjusted Rand Index의 수렴여부의 예



〈그림 5〉 시뮬레이션결과에 따른 평균 Adjusted Rand Index의 히스토그램의 예

임을 보여주고 있다. 누락된 속성의 수가 1개, 그리고 2개일 때는 상관계수에 의한 방법(방법 I)이 좋은 예측을 보여주었으며, 누락된 속성의 수가 증가하여 3개가 되었을 경우에는 Correspondence Analysis를 이용한 자료융합 방법(방법 II)이 더 좋은 예측을 보여주고 있다. Adjusted Rand Index를 이용한 자료융합 방법의 평가에 대한 결과도 상관 계

〈표 5〉 Monte Carlo Simulation에 의한 Adjusted Rand Index의 분포특성
(누락된 속성의 수 기준)

	집단 A 삭제된 속성의 수 = 1		집단 B 삭제된 속성의 수 = 2		집단 C 삭제된 속성의 수 = 3	
	방법 I	방법 II	방법 I	방법 II	방법 I	방법 II
평 균	0.5605	0.5300	0.3100	0.2809	0.1657	0.1888
표준편차	0.1085	0.1092	0.1140	0.1214	0.0820	0.0840

〈표 6〉 Monte Carlo Simulation에 의한 Adjusted Rand Index의 분포특성
(군집의 수 기준)

	군집의 수 = 2		군집의 수 = 3		군집의 수 = 4		군집의 수 = 5	
	방법 I	방법 II	방법 I	방법 II	방법 I	방법 II	방법 I	방법 II
평 균	0.4421	0.4448	0.3423	0.3165	0.3064	0.2960	0.2907	0.2755
표준편차	0.1492	0.1450	0.1984	0.1832	0.2038	0.1820	0.1782	0.1496

〈표 7〉 Monte Carlo Simulation에 대한 ARI의 분산분석 결과

Source	d.f.	Sum of square	Mean square	F value	Pr > F
Model	17	68.8268	4.0486	693.56	< 0.0001
방 법	1	0.08866	0.08866	15.19	0.0001
군집의 수	3	9.3361	3.1120	533.11	< 0.0001
누락된 속성의 수	2	56.4783	28.2391	4837.53	< 0.0001
방법 * 군집의 수	3	0.0627	0.0209	3.58	0.0133
방법 * 속성의 수	2	0.3730	0.1865	31.95	< 0.0001
Error	2382	13.9049	0.0058	-	-
Total	2399	82.7317	-	-	-

수를 이용한 평가에서 얻은 결과와 일관성 있는 결과를 관측할 수 있었다. 또한 방법 II가 방법 I에 비하여 평균 Adjusted Rand Index의 분포에 대한 편차가 크다는 현상도 두 가지 방법에 의한 평가에서 모두 관측되었다.

<표 6>는 군집의 수에 따라 Adjusted Rand Index가 어떻게 변화하는지를 보여 주고 있다. 예상했던 바와 같이 군집의 수가 증가할수록 두 가지 자료융합 방법에 의한 본래 군집 구조의 도출 성과가 현저히 저하됨을 관측할 수 있다. 그리고, 군집의 수가 증가할수록 상관계수에 의한 방법(방법 I)이 Correspondence Analysis를 이용한 자료융합 방법(방법 II)보다 더 좋은 예측을 보여주고 있다.

<표 7>은 Adjusted Rand Index의 분산분석 결과를 보여 주고 있다. 두 가지의 자료융합 방법, 군집의 수, 그리고 누락된 속성의 수에 따라 Adjusted Rand Index의 값이 유의적으로 차이가 있음을 볼 수 있다. 즉, 군집의 수가 증가할수록, 그리고 누락된 속성의 수가 증가할수록 예측된 값의 일치도는 감소함을 볼 수 있다. 또한 그 일치도는 상관계수를 사용한 자료융합과 Correspondence Analysis를 사용한 자료융합 방법에 따라 발생한 차이가 통계적으로 유의한 결과가 나타나 있다. 누락된 속성의 수와 두 가지 다른 융합방법, 군집의 수와 두 가지 다른 융합방법은 교호효과(interaction effect)가 존재하고 있음을 보여주고 있다.

5. 결론 및 향후 연구방향

본 연구에서는 자료융합 방법을 누락치의 추정에 적용하여 그 성과를 실제자료의 시뮬레이션을 통하여 탐색적으로 평가하였다. 두 가지의 누락치 추정방법과 두 가지의 평가기준을 채택한 연구결과는 전반적으로 고무적이라고 할 수 있다. 상관계수와 군집의 안정도(Adjusted Rand Index)를 평가기준으로 사용한 경우 모두 누락된 속성의 수가 증가함에 따라 방법 II(Correspondence Analysis에 의한 자료융합)가 방법 I(상관계수를 이용한 자료융합) 보다 성과가 우수하게 나타났다. 그러나, 군집의 안정도를 이용한 평가에서는 군집의 수를 증

가 시킴으로 해서 방법 I(상관계수를 이용한 자료융합)이 방법 II(Correspondence Analysis에 의한 자료융합) 보다 성과가 우수하게 나타났다.

본 연구는 탐색적인 연구의 성격에도 불구하고 향후 연구를 위한 다음과 같은 문제를 제기하고 있다.

- 본 연구에서는 13개의 속성 중 임의로 처음의 6개를 공통변수로 선정하였다. 공통변수의 선택은 각각의 수혜자에 대한 기증자를 선택하는 근거가 된다는 점에서 매우 중요하다. 향후 연구는 공통자료의 선택에 관하여 진행되어야 할 것이다. 가능한 공통변수로서는 응답자의 이상점(Ideal Point), 속성의 중요도(Attribute Importance Weight), 혹은 응답자의 배경변수(인구통계적 변수, 사이코 그래픽 변수, 라이프스타일 등을 포함) 등을 사용할 수 있다. 또한 이들 공통변수의 척도에 따라 상관계수(product moment correlation) 혹은 등급상관계수(rank correlation)를 누락치의 추정에 적용할 수 있을 것이다. 향후 연구에서는 이러한 공통변수의 선택이 추정치의 정확도에 미치는 영향을 실펴 보아야 할 것이다.
- 향후연구에서는 또한 군집을 도출하는데 사용한 프로그램의 비교를 다루어야 할 것이다. 특히 K-평균 군집분석은 그 초기치의 선정에 매우 민감한 결과를 나타내므로 군집의 안정도를 평가기준으로 채택하는 경우 군집분석 프로그램의 선정에 주의하여야 한다(Berry & Linoff, 1997 ; Helsen & Green, 1991). 이러한 초기치의 선택에 의한 영향을 최소화하기 위하여 Convergent Cluster Analysis(CCA) 이나 CONCLUS Program 등이 Sawtooth Software사에 의하여 개발되어 쓰이고 있는데, 이는 군집분석의 반복을 통한 최적 군집의 도출을 하고 있다(Helsen & Green, 1991). 그 밖에도 Berry & Linoff(1997)는 여러 가지의 초기치 결정방법과 군집간의 계산방법을 통하여 K-평균 군집방법의 정확도를 향상시키는 시도를 한 바 있다. 본 연구에서는 컨조인트 부분가치자료를 이용하였으나 향후 연구에서는 다른 종류(응답자의 매체행동, 태도, 브랜드선호도 등)의 실제자료를 사용하는 시도가 필요할 것이다.

본 연구에서는 두 가지 다른 누락치 추정 방법들(상관계수를 이용한 자료융합과 Correspondence Analysis를 이용한 자료융합)의 성과를 다양한 환경 하에서 비교하는데 주요 목적이 있다. 실무에서는 누락된 고객 데이터를 추정하기 위하여 다양한 방법을 사용하고 있다. 그러나, 아직까지 누락된 데이터를 추정하는 방법들 간의 성과 비교에 관한 연구는 없었다. 본 연구에서의 결과는 크게 두 가지 방면에서 효율적으로 고객 데이터를 관리할 수 있도록 하는데 도움을 줄 수가 있다. 첫 번째로는, 고객 분석을 위하여 실시한 설문지 조사에서 응답자의 불성실한 답변으로 누락된 문항에 대한 답을 추정하는데 본 연구에서 제시한 방안을 사용할 수 있을 것이다. 모든 설문지 문항이 응답자에 의해 성실히 답변 되어진 자료만을 사용하여 고객을 분석한다면 많은 수의 응답자로부터 의견을 수집하기 위해서는 많은 비용과 노력이 소요될 것이다. 본 연구의 결과를 기초로 하여 체계적으로 누락된 값을 추정한다면 정보 수집의 생산성을 높일 수 있을 것이다. 두 번째로는, 많은 설문지 문항 수로 인한 응답자의 불성실한 답변을 방지하기 위한 하나의 방안으로 본 연구의 결과를 활용할 수 있을 것이다. 설문지 디자인 단계부터 고객의 다양한 면(예를 들어, 자동차에 대한 선호도 조사, 컴퓨터에 관한 선호도 조사, 커피에 대한 선호도 조사 등)에 대한 여러 개의 설문지를 디자인하여 본 연구에서 제시한 자료융합 방법을 사용하여 여러 설문지 조사로부터 얻어진 자료를 융합한다면 정보 수집의 생산성은 크게 높아질 수 있을 것이다.

참 고 문 헌

[1] Baker, K., P. Harris, and J. O'Brien, "Data Fusion : An Appraisal and Experimental Evaluation," Journal of the Market Research Society, 39(1), (1997), pp.227-271.

[2] Berry, M. and G., Linoff, Data Mining Techniques, Wiley Computer Publishing, 1997.

[3] Carroll, J.D. and P. Arabie, "Multidimensional Scaling," in M. R. Rosenzweig and L. W. Porter (eds.), Annual Review of Psychology, Volume 31, Palo Alto, CA : Annual Review, (1980), pp.607-49.

[4] Carroll, J.D., P.E. Green, and C.M. Schaffer, "Comparing Interpoint Distances in Correspondence Analysis," Journal of Marketing Research, 24, (November 1987), pp.455-50.

[5] Carroll, J.D., P.E. Green, and C.M. Schaffer, "Interpoint Distance Comparisons in Correspondence Analysis," Journal of Marketing Research, 23 (August 1986), pp.271-80.

[6] Downey, R.G. & C.V. King, "Missing Data in Likert Ratings : A Comparison of Replacement Methods," The Journal of General Psychology, 125(2), (1998), pp.175-191.

[7] Helsen, K. and P.E. Green, "A Computational Study of Replicated Clustering with an Application to Market Segmentation," Decision Sciences, 22, (1991), pp.1124-1141.

[8] Hoffman, D.L. and G.R. Franks, "Correspondence Analysis : Graphical Representation of Categorical Data in Marketing Research," Journal of Marketing Research, 23 (August 1986), pp.213-27.

[9] Kamakura, W.A. & M. Wedel, "Factor Analysis and Missing Data, Journal of Marketing Research," Journal of Marketing Research, (2000), pp.490-498.

[10] Kromrey, J.D. & C.V. Hines, "Nonrandomly Missing Data in Multiple Regression : An Empirical Comparison of Common Missing-Data Treatments," Educational and Psychological Measurement, 54(3), (1994), pp.573-593.

[11] Kruskal, J. and M. Wish, Multidimentional Scaling, Newbury Park, CA., 1978.

[12] Landerman, L.R., K.C., Land, and Pieper, C.F., "An Empirical Evaluation of the Predictive Mean matching Method for Imputing Missing Values," Sociological Methods & Research, 26(1), (1997), pp.3-33.