마할라노비스 거리를 이용한 자료융합전략의 성과측정*

김성호
한양대학교 경영대학 교수
(kim007@hanyang.ac.kr)
조성빈
서강대학교 경영대학 교수
(sungbincho@sogang.ac.kr)

본 연구는 자료융합에서 효과적인 데이터베이스의 축소와 설계에 기여할 수 있는 여러 가지 기증자선택전략을 탐구하고 그 성과를 비교하고자 한다. 공통속성변수를 미리 정하여 놓은 계획된 자료융합을 설정하였으며, 자료융합의 영역에 최초로 마할라노비스 거리의 개념을 도입하고 응답자간의 비유사성을 측정하였다. 본 연구에서는 다섯 가지의 기증자선택전략이 설계되었다: 상관계수를 적용하는 전략; 유클리디안 거리를 적용하는 전략; 마할라노비스 거리를 적용하는 전략; 대응일치분석을 거쳐 유클리디안 거리를 적용하는 전략; 대응일치분석을 거쳐 마할라노비스 거리를 적용하는 전략. 몬테카를로 시뮬레이션을 실시한 결과, 대응일치분석을 거쳐 마할라노비스 거리를 적용하는 전략의 성과가 가장 우수하게 도출되었으며, 다음으로는 대응일치분석을 거쳐 유클리디안 거리를 적용하는 전략으로 나타났다. 본 연구의 실험설계와 결과에서 발견한 점으로는 첫째, 대응일치분석은 공통속성변수의 차원을 축소하고 보다 변별력을 향상시키는데 기여하였으며, 둘째, 마할라노비스 거리의 도입은 공통속성변수 간에 실재하는 통계적 종속성을 모델에 반영함으로써 예측의 성과를 높이는데 유효한 역할을 한 것으로 판단된다.

Ⅰ. 서론

급변하는 경영환경과 치열한 경쟁에서 생존하기 위하여 많은 기업들은 고객과 시장의 니즈(needs)에 신속하고 효과적으로 반응하여야 할 필요성이 커지고 있다. 대부분의 대기업들은 이미 고객관계관리(customer relationship management: CRM)를 기업성공에 꼭 필요한 주요 전략으로 삼았으며, 근래에는 더 나아가 e-CRM 구축에 촉각을 곤두세우고 있다. e-CRM은 고객 정보의 수집과 활용에 있어 인터넷을 기반으로 하여 고객이 인식하지 못하는 차원의 데이터까지도 수집하여 고객의 모든 정보와 성향을 실시간으로 분석하고 마케팅활동으로 바로 연결이 가능한 솔루션이라고 할 수 있다. 인터넷 기반의 고객관계관리는 대기업뿐만 아니라 중소기업에서도 비교적 적은 비용으로 고객관계관리 전략을 구사할 수 있도록 할 수 있으리라 기대된다. 고객 니즈를 정확히 파악하는 것이 중요하다는 인식이 확산되면서 기업들은 온라인과 오프라인의 모든 채널을 동원하여 고객을 만족시키기 위하여 많은 자금과 노력을 투자하고 있다. 성공적인 고객관계관리의 이행은 고객 데이터를 적시에 수집하고 기업의 성과향상에 이용함으로써 가능할 것이다.

재화와 서비스 측면에서 고객의 행위를 분석하는 기법중의 하나로서 컨조인트 분석기법을 들 수 있다. 컨조인트 분석기법의 광범위한 적용성은 잘 알려져 있지만 그에 못지않게 평가하여야 할 가상 제품의 수가 기하급수적으로 늘어나는 단점을 가진

논문접수일: 2005. 5 게재확정일: 2005. 9
* 본 연구는 2003년도 한양대학교 교내연구비에 의하여 수행되었음.

것으로도 유명하다. 가장 빠르고 경제적인 고객 데이터의 수집은 설문을 통하여 할 수 있다. 설문을 어떻게 디자인하고 분배하며 회수하여 정리하는 가에 관하여 여러 가지 연구가 진행되어 왔는데 연구에 비하여 결과는 그리 성공적이지 못하여 왔다고 할 수 있다. 요즈음과 같이 인터넷을 통한 설문이 보편화되어가고 있는 상황에서, 응답자들이 설문에 답하는 도중 주위 상황에 이끌려 주의가 산만하여지거나 일정한 정도의 집중을 기울여 설문을 완성하지 못하는 것이 일반적이다. 따라서 응답자가 고도의 주의를 기울여 성실하게 모든 문항에 답변할 수 있도록 문항을 개발하고 답하여야 할 문항의 수나 양도 줄이는 것이 바람직하다. 그렇게 함으로써 양질의 고객 데이터를 수집하고 고객관계관리 전략을 성공적으로 이행할 수 있을 것이다. 본 연구는 고객의 정보를 수집하고 관리할 수 있는 방법론중 하나인 자료융합 기법을 소개하고, 기증자선택 전략 중의 하나로서 마할라노비스 거리(Mahalanobis distance)를 이용한 전략을 자료융합 연구에 최초로 시도하고 그 성과를 측정하는데 주요한 목적을 두고 있다.

II. 컨조인트 분석기법

컨조인트 분석기법(conjoint analysis)은 수리심리학에서 유래된 기법으로서 고객의 재화와 서비스에 대한 선호도발달에 관한 이해를 돕는 도구로 사용되고 있다. 선호도는 다속성(multi-attributes)과 다속성수준(multi-attribute levels)으로 표현된다. 효용(utility)은 각각의 속성수준에 대한 고객의 주관적 선호를 측정하는 수단으로 볼 수 있다. 각 속성수준에 대한 효용은 모두 합산되어 총효용을 나타낸다. 그리하여 총효용이 큰 재화나 서비스가 여타의 재화나 서비스에 비하여 고객에게 좀 더 나은 선택으로 간주되어진다. 컨조인트 분석기법의 강점은 계량형과 비계량형 종속변수를 모두 다룰 수 있다는 점과 독립변수와 종속변수에 대한 일반적 가정에서 찾을 수 있다(Hair et al., 1998). Green and Rao(1971)에 의하여 의사결정문제와 마케팅문제의 해결에 도입된 이후로, 컨조인트 분석기법은 지난 30여 년간 구매자의 상품선호도를 분석하는 도구로 널리 애용되어 왔다(Green and Srinivasan, 1990). 그 외에도 시장세분화와 최적의 상품 포지셔닝(positioning)에도 사용되어 왔다(Green and Krieger, 1993).

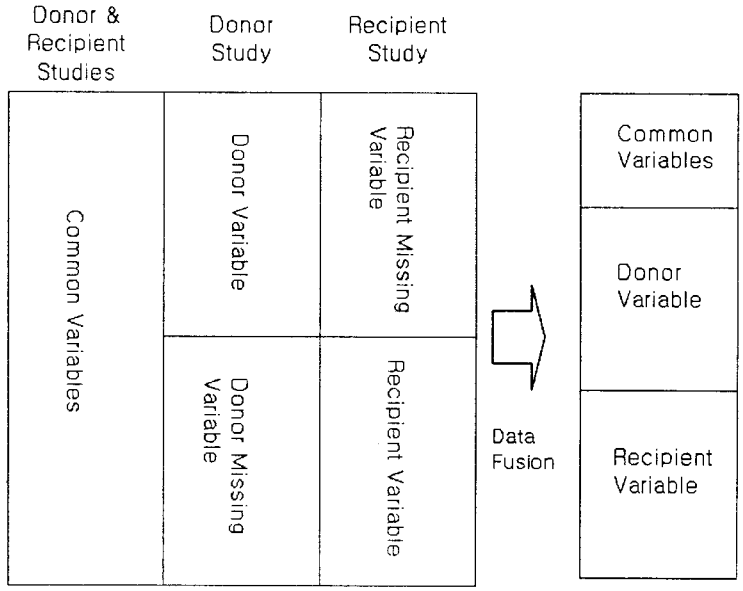
그럼에도 불구하고 컨조인트 실험에 참가하는 응답자들이 수많은 가상제품들을 비교하는 것이 단점으로 지적되어 왔다(Wittink and Cattin, 1989). 속성과 속성수준의 수가 늘어남에 따라 비교하여야 할 가상제품의 수는 기하급수적으로 증가하게 된다. 통상 하나의 제품 평가에 10개 이상의 속성을 포함시킨다는 것을 감안한다면, 하나의 속성에 3개씩의 속성수준만을 가져도 응답자는 3^{10}개의 가상제품을 비교하여야 하는 큰 어려움에 직면하게 되는 셈이다. 전통적으로 fractional factorial design을 사용하여 비교 가상제품의 수를 줄여 왔는데, Kim and Hamano(1995)가 자료융합 기법을 컨조인트 분석기법에 새로이 적용하여 그 수를 한층 감소시키려고 시도한 바 있다. 용인될 만한 수준의 정확도에서 누락치에 대한 대체값으로 자료융합 기법이 성공적으로 사용될 수도 있음을 보여 주었다.

Ⅲ. 자료융합

자료융합(data fusion)은 두 개 이상의 표본집단으로부터의 서로 다른 설문지 조사 자료를 융합하여 하나의 융합되어진 자료를 만들어 내는 과정이다(Kamakura and Wedel, 2000). 자료융합은 커뮤니케이션 분야에서 뿐 만 아니라 설문지를 통한 마케팅 조사 등에서 빈번하게 발생하는 누락치를 추정하는 데에도 유용하게 사용될 수 있다. 자료융합은 크게 세 가지 분야에서 널리 활용되어지고 있다. 첫째 활용분야는 시장조사이다. 여러 가지 자료를 융합하여 하나의 필요한 자료를 만들어 낸다. 분야가 서로 다르더라도 공통되거나 유사한 변수에 기초하여 자료융합 방법을 사용함으로써, 직접 조사하지 않더라도 그 표본 집단의 의견을 유추할 수 있다. 둘째는 여러 가지 종류의 제품에 대한 선호도 조사 자료를 융합하여 고객이 하나의 제품에 대하여 직접 설문지에 답하지 않았을 지라도 그 제품에 대한 다른 응답자의 선호도를 기초로 하여 그 고객의 선호 정도를 예측할 수 있다. 다시 말하면, 자료융합을 통하여 추정된 고객의 선호도를 기초로 하여 선별적으로 고객에게 direct marketing을 실시함으로 해서 그것의 효율성을 높일 수 있다. 셋째는 미디어 계획에 쓰일 수 있다. 고객에 대한 구매행동에 관한 자료와 TV 프로그램에 대한 시청률 조사 자료를 융합함으로써 광고나 판촉 기획의 기본 자료로 많이 사용하고 있다.

자료융합에서는 응답자 표본을 기증자(donor)와 수혜자(recipient)로 구분한다. 기증자란 어떤 특정한 수혜자의 누락치에 대한 대체값을 제공해 주는 응답자를 말하며, 수혜자란 어떤 특정한 기증자로부터 누락치에 대한 추정값을 제공받는 응답자를 말한다. 〈그림 1〉은 자료융합 과정을 설명해 주고

〈그림 1〉 자료융합 과정



있다. 자료융합에서 각각의 수혜자는 속성에 기초하여 자신과 가장 유사한 기증자를 찾는다. 유사하다는 의미는 두 가지로 해석할 수 있다. 한가지는 속성상 유사함이 큰 기증자를 찾는다는 것으로 해석할 수 있으며, 또 다른 한가지는 유사성의 정의가 모호하고 힘들다는 점을 감안하여 비유사성을 측정하고 비유사성이 가장 작은 기증자를 찾는 방법이다.

일반적으로 공통자료에는 응답자의 인구 통계적 정보(demographic information)가 포함되어 왔다. 일단 특정 수혜자가 자신과 가장 유사한 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 유사한 응답자를 수혜자로 찾는다. 수혜자와 기증자와의 유사성을 정의하기 위하여 상관계수(correlation coefficient)가 사용되며, 비유사성의 정의에는 거리개념(distance measure)이 사용된다. 만약 수혜자나 기증자의 비누락치(non-missing value)로 구성된 공통자료가 범주형 변수(categorical variables)라면, 어떤 특정한 차원을 지닌 공간 내에서 각각의 수혜자에 대한 기증자를 파악하기 위하여 응답자간의 거리 혹은 상관계수를 직접적으로 계산하는 것은 불가능하다. 이 때 응답자간의 거리를 계산하기 위하여 대응일치분석을 사용할 수 있다. 기존의 연구에서는 응답자의 실수와 부주의 혹은 피곤 등으로 인한 비의도적인 자료의 누락치(unintentional data missing)를 예측하는 방법에 대한 성과를 비교하였다(Kromrey and Hines, 1994; Landerman et al., 1997). 그러나, 공통변수를 미리 정해놓은 의도적인 누락치(intentional data missing)의 대체값을 제공해 주는 방법들에 대한 연구는 드물다(Baker et al., 1997). 본 연구에서는 의도적이며 계획된 자료융합에서 기증자 선택을 어떻게 할 것인가에 관한 여러 가지 시도를 하여 보고자 한다.

Ⅳ. 대응일치분석

대응일치분석(correspondence analysis)이란 명목척도(nominal scale)로 측정된 범주형 변수의 관계를 분석하는 방법으로 두 변수의 각 범주들을 공간(일반적으로 2차원공간)상에 점으로 표현하여 범주들 사이의 관계를 분석하는 방법이다(Hoffman and Franke 1986; Carroll et al., 1987). 이와 같이 범주들을 공간상에 점으로 표현함으로서 범주들 사이의 유사성여부를 시각적으로 관측할 수 있다. 이 분석방법에서 분할표의 각 칸(cell)에 주어진 자료는 두 변수사이의 관계를 측정하는 어떠한 값도 가능한데, 일반적으로 관측도수를 이용하는 분할표 자료가 많이 이용된다. 대응일치분석은 입력자료가 dummy 변수를 포함한 명목 척도이거나 혹은 응답 항목(dichotomous이거나 multichotomous)에 따른 응답의 빈도수(frequency)라는 점에서 일반적인 다차원 척도법과 구분되어진다. 따라서 대응일치분석은 일반적으로 n-way contingency table 혹은 cross-tabulation을 분석하는데 사용되는 기법이다.

다른 다차원 척도법과 마찬가지로 입력자료에 나타난 개체(예를 들어, 소비자, 제품, 기업, 제품의 사용상황 등)들을 일반적으로 p-차원의 공간에 점(point)으로 나타내는 기법이다. 이 과정에서 입력자료에 나타난 개체들간의 상대적인 관계를 대응일

치분석을 통하여 구성한 공간에서도 동일하게 유지하여 나타낸다는 것이 특징이라고 할 수 있다. 또한 입력자료의 가로와 세로에 나타난 개체들을 동시에 동일한 공간에 점으로 나타내는 Joint Space 분석기법의 일종이라고 할 수 있다. 또한 대응일치분석의 알고리즘에 따라서는 가로와 세로의 개체들 간의 거리가 직접 비교가능 할 수도 있다.

Carroll et al.(1986)의 알고리즘에 따르면, 우선 분석의 대상이 되는 contingency table 혹은 cross tabulation matrix F 행렬(F는 가로 I 줄과 세로 J 줄로 되어 있다고 하자)을 다음과 같이 H 행렬로 정규화(normalize)한다.

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

R 행렬은 $I \times I$ diagonal matrix이며 C 행렬는 $J \times J$ 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 H 행렬은 다음의 식을 사용하여 chi-square 거리척도로 전환된다.

$$H = P \Delta Q'$$

여기에서 $P'P = Q'Q = I$이며 Δ는 diagonal metric이다. 끝으로 p-차원 상에서의 가로줄(X)과 세로줄(Y)에 나타난 개체의 좌표는 각각 다음과 같다.

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

V. 연구방법

5.1 연구설계

자료융합은 앞에서 언급한 바와 같이 의도하지 않은 자료융합과 의도된 자료융합으로 구분할 수 있다. 의도하지 않은 자료융합은 사후적(ex post) 자료융합으로서 두 개 이상의 데이터 베이스를 융합 할 때 존재하는 누락치에 대한 대체값을 추정하는 경우이다. 의도된 자료융합이란 사전적으로(ex ante) 계획된 자료융합(preplanned data fusion)으로서, 두 개 이상의 데이터 베이스를 구축할 때, 차후에 자료융합 할 것을 의도하여 계획적으로 축소된 데이터 베이스를 설계하는 것이다. 본 연구에서는 계획된 자료융합 방법을 채택하고 그 안에서 기증자 선택에 관한 전략을 살펴보고자 한다. 특히 자료융합 최초로 마할라노비스 거리(Mahalanobis distance) 개념을 도입하여 응답자간의 비유사성의 척도로 사용하고 유용성을 평가하고자 한다. 본 연구의 구체적인 목적은 다음과 같다.

- 계획된 자료융합에서 다양한 기증자선택전략들의 성과를 비교한다.
- 누락된 속성의 양이 추정치의 정확도에 미치는 영향을 분석한다.

연구에 사용된 자료는 신용카드의 속성과 속성수준에 대한 자료로서, 〈표 1〉에 요약된 것처럼 12개의 속성으로 이루어져 있고 각 속성은 다수의 속성수준으로 구성되어 총 35개의 속성수준이 존재한다. 각 속성수준에 대한 응답자 480명의 부분효용(part-worth)이 측정되었다.

〈표 1〉 신용카드 자료의 속성과 속성수준

속성	속성수준
연회비	$0; $10; $20; $50; $80; $100
현금환불	None; 1/2%; 1%
메시지전달 무료전화서비스	None; 9-5 PM weekdays; 24 hours day
구매상품보험	None; 90 days coverage
항공기탑승보험	None; $50,000; $200,000
렌털카보험	None; $30,000
수하물보험	None; $2,500 deprecated cost; $2,500 replacement cost
공항클럽/라운지 이용	No admission; $5/visit; $2/visit
신용카드 사용가능 장소	Air, hotel, rental car (AHC); AHC and restaurant (AHCR); AHCR and most general retailers; AHCR and department stores only
비상차량지원 무료전화서비스	No; yes
공항-시내간 리무진 서비스	Not offered; available at 20% discount
24시간 의료/.법률상담 소개서비스	Not offered; available at 20% discount

Kim et al. (2004)의 연구에 따르면, 공통변수의 선택에서 무작위로 하는 것 보다 체계적인 방법으로 variance 혹은 weight를 사용하는 전략이 자료융합의 성과를 향상시킨 바 있다. 따라서 본 연구에서는 12가지의 속성에서 누락되지 않은 공통변수(common variable)의 역할을 담당할 여섯 가지의 속성을 선별하는 기준으로는 분산(variance)의 크기를 채택하였다. 각 속성변수에 대한 분산은 속성내 분산(within-attribute variance)과 속성간 분산(between-attribute variance)의 합에 의하여 구하여진다.

$$Var[Y_k] = E[Var(Y_{kl} \mid L)] + Var[E(Y_{kl} \mid L)]$$

(Y_k: 속성 k 변수

Y_{kl}: 속성 k의 속성수준 l 변수

L: 속성수준변수의 종류)

무작위로 공통변수를 선택하는 것보다 자료융합의 정확도를 향상시키기 위하여 분산의 크기가 큰 여섯 가지 속성인 연회비, 현금환불, 신용카드 사용가능 장소, 비상차량지원 무료전화서비스, 공항-시내간 리무진 서비스, 24시간 의료/법률상담 소개서비스가 공통속성으로 채택되었다. 공통변수로 채택되지 아니한 나머지 여섯 가지의 속성은 비공통변수(non-common variable)로서 데이터 베이스의 구축 시에 수집되지 아니할 속성으로 채택되었다.

비공통변수에 대하여는 삭제속성의 수를 1개, 2개, 3개로 증가시켜 보면서 자료융합의 정확도에 미치는 영향을 분석하였다. 구체적인 절차는 다음과 같다.

1. 응답자 480명을 두 집단으로 나눈다 ($n_1 = n_2 = 240$).

2. 집단 1(n_1)에서 비공통변수로 선택된 속성 중 무작위로 하나의 속성을 선택하고 그에 속한 모든 속성수준을 누락시킨다.
3. 집단 2(n_2)에서 비공통변수 중 집단 1에서 누락된 속성을 제외한 나머지의 속성 중 하나를 무작위로 선택하여 그에 속한 모든 속성수준을 누락시킨다.

두 집단간에 상이한 속성을 누락시킴으로써, 두 개의 데이터 베이스를 구축할 때 누락된 속성은 수집하지 아니한 것으로 간주할 수 있다. 다시 말하면, 계획된 자료융합을 통하여 축소된 데이터 베이스를 구현하는 것을 모델하였다. 삭제속성의 수는 다음과 같이 증가시키며 효과를 측정하였다.

- 삭제 경우 1: 두 집단에 대하여 비공통변수 중 서로 상이한 한 가지의 속성을 무작위로 선택하여 그에 해당되는 속성수준모두를 누락시킨다.
- 삭제 경우 2: 두 집단에 대하여 비공통변수 중 서로 상이한 두 가지의 속성을 무작위로 선택하여 그에 해당되는 속성수준모두를 누락시킨다.
- 삭제 경우 3: 두 집단에 대하여 비공통변수 중 서로 상이한 세 가지의 속성을 무작위로 선택하여 그에 해당되는 속성수준모두를 누락시킨다.

본 연구에서는 기증자선택전략으로 다음에 소개하는 다섯 가지 방법을 적용하였다.

전략 1

상관계수(correlation coefficient)를 이용한다. 공통변수로 선택된 여섯 가지의 속성에 속하는 모든 속성수준 즉, 19개의 속성수준을 사용하여 응답자간의 상관계수를 계산한다. 응답자 i와 j의 상관계수는 다음과 같다($i \neq j$):

$$r_{ij} = \frac{\sum_{all\ kl} (\bar{Y}_{i,kl} - Y_{i,kl})(\bar{Y}_{j,kl} - Y_{j,kl})}{\sqrt{\sum_{all\ kl} (\bar{Y}_{i,kl} - Y_{i,kl})^2} \sqrt{\sum_{all\ kl} (\bar{Y}_{j,kl} - Y_{j,kl})^2}}$$

($Y_{i,kl}$: 응답자 i의 속성 k 속성수준 l 변수)

집단 1의 응답자가 가진 누락치는 집단 2에 속한 응답자 중 상관계수가 가장 큰 응답자의 속성수준으로 대체한다. 반대로 집단 2의 응답자에 대한 누락치는 집단 1에서 상관계수가 가장 큰 응답자를 찾아 대체값으로 이용한다.

전략 2

유클리디안 거리(Euclidean distance)를 이용한다. 공통변수로 선택된 여섯 가지의 속성에 속하는 모든 19개의 속성수준을 사용하여 응답자간의 거리를 구한다. 응답자 i와 j의 유클리디안 거리를 구하는 공식은 다음과 같다($i \neq j$):

$$ed_{ij} = \sqrt{(\tilde{Y}_{i,kl} - \tilde{Y}_{j,kl})(\tilde{Y}_{i,kl} - \tilde{Y}_{j,kl})^T}$$

($\tilde{Y}_{i,kl}$: 응답자 i의 속성 k 속성수준 l 변수 벡터)

집단 1의 응답자가 가진 누락치는 집단 2에 속한 응답자 중 유클리디안 거리가 가장 작은 응답자의 속성수준으로 대체한다. 반대로 집단 2의 응답자에 대한 누락치는 집단 1에서 유클리디안 거리가 가장 작은 응답자를 찾아 대체값으로 이용한다.

전략 3

마할라노비스 거리(Mahalanobis distance)를 이용한다. 공통변수로 선택된 여섯 가지의 속성에 속하는 모든 19개의 속성수준을 사용하여 응답자간의 거리를 구한다. 응답자 i와 j의 마할라노비스 거리를 구하는 공식은 다음과 같다($i \neq j$):

$$md_{ij} = \sqrt{(\tilde{Y}_{i,kl} - \tilde{Y}_{j,kl})\Gamma^{-1}(\tilde{Y}_{i,kl} - \tilde{Y}_{j,kl})^T}$$

(Γ^{-1} : 속성수준에 대한 kl by kl (19 차원) 공분산 메트릭스)

집단 1의 응답자가 가진 누락치는 집단 2에 속한 응답자 중 마할라노비스 거리가 가장 작은 응답자의 속성수준으로 대체한다. 반대로 집단 2의 응답자에 대한 누락치는 집단 1에서 마할라노비스 거리가 가장 작은 응답자를 찾아 대체값으로 이용한다.

전략 4

공통변수에 대하여, 각 속성수준 중 가장 큰 값을 갖는 속성수준의 종류를 그 속성수준이 속한 속성의 이상점(ideal point)으로 인식하여 범주형 변수(categorical variable)로 취급한다(김성호와 이경미, 1999). 6개로 축소된 범주형 공통변수에 대하여 SAS의 대응일치분석을 적용하여 각 응답자의 5차원 좌표(공통속성의 수보다 하나 적은 차원으로 축소됨)를 구한 후, 이 좌표를 이용하여 응답자간의 유클리디안 거리를 구한다.

$$ced_{ij} = \sqrt{(\tilde{d}_{i,k} - \tilde{d}_{j,k})(\tilde{d}_{i,k} - \tilde{d}_{j,k})^T}$$

($\tilde{d}_{i,k}$: 대응일치분석에서 얻어진 응답자 i의 k 차원 좌표 벡터)

누락치에 대한 대체값을 찾는 방법은 전략 2와 같다.

전략 5

전략 4와 마찬가지로 공통변수에 대하여 각 속성의 범주형 이상점을 구하고, 6개로 축소된 범주형 공통변수에 대하여 SAS의 대응일치분석을 통하여 응답자의 5차원 좌표를 구한 후, 이 좌표를 이용하여 응답자간의 마할라노비스 거리를 구한다.

$$cmd_{ij} = \sqrt{(\tilde{d}_{i,k} - \tilde{d}_{j,k})\Gamma^{-1}(\tilde{d}_{i,k} - \tilde{d}_{j,k})^T}$$

(Γ^{-1} : 속성 k 차원(5 차원)에 대한 공분산 메트릭스)

누락치에 대한 대체값을 찾는 방법은 전략 3과 같다.

5.2 연구 결과

본 연구에서는 기증자선택전략 다섯 가지, 누락 정도 세 가지 별로 20회의 몬테카를로 시뮬레이션(Monte Carlo simulation)을 실시하였다(실험 횟수: 5 × 3 × 20). 데이터를 누락시키기 전의 본래 데이터 베이스와 누락시킨 후 자료융합을 이용하여 대체값을 추정하여 완성시킨 데이터 베이스가 얼마나 근사한가를 평가하기 위하여 두 가지 척도를 사용하였다. 두 데이터 베이스간의 상관계수와 root mean squared error(root MSE)를 산출하였다. 첫째 평가척도인 상관계수에 대한 분산분석의 결과는 〈표 2〉에 요약되어있다. 전략요인과 삭제요인 모두 통계학적으로 유의하게 나타났으며, 전략과 삭제에 의한 교호효과도 존재하는 것으로

〈표 2〉 본래의 데이터 베이스와 자료융합된 데이터 베이스의 상관계수에 대한 분산분석결과

Source of variation	Sum of squares	d. f	Mean squares	F value	p-value
모델	0.2115	14	0.0151	469.41	〈0.0001
전략	0.0035	4	0.0008	27.03	〈0.0001
삭제 경우	0.2071	2	0.1036	3217.58	〈0.0001
전략 * 삭제 경우	0.0009	8	0.0001	3.56	0.0006
오차	0.0092	285	0.00003		
합계	0.2207	299			

〈표 3〉 본래의 데이터 베이스와 자료융합된 데이터 베이스의 root MSE에 대한 분산분석결과

Source of variation	Sum of squares	d. f	Mean squares	F value	p-value
모델	0.001402	14	0.000100	500.59	〈0.0001
전략	0.000013	4	0.000003	16.35	〈0.0001
삭제 경우	0.001386	2	0.000693	3464.38	〈0.0001
전략 * 삭제 경우	0.000002	8	0.000000	1.77	0.0827
오차	0.000057	285	0.000000		
합계	0.001459	299			

나타났다. 〈표 3〉은 또 다른 평가척도인 root MSE를 적용하여 분산분석을 실시한 결과로서 전략요인과 삭제요인이 유의한 영향을 미치는 것으로 나타났다.

〈표 4〉와 〈표 5〉는 실험 요인에 대한 각각 평균 상관계수와 평균 root MSE를 보여주고 있다. 첫 번째 발견점은 비공통변수에 대한 삭제의 수가 증가할수록 자료융합의 정확도가 낮아지고 있음이다. 평균 상관계수를 보면, 삭제 경우 1은 0.9662, 삭제 경우 2는 0.9327, 삭제 경우 3은 0.9018로 나타나고 있다. 평균 root MSE는 삭제 경우가 증가함에 따라 0.005263, 0.008329, 0.010503으로 증가하고 있다.

둘째로 전략에 의한 비교에서는, 대응일치분석을 거쳐 마할라노비스 거리를 이용한 전략 5에 의한 자료융합의 성과가 가장 좋은 것으로 나타났다. 그 다음으로는 대응일치분석을 거쳐 유클리디안 거리에 기초한 전략의 정확도가 높은 것으로 나타났고, 상관계수에 의한 전략, 유클리디안 거리에 의한 전략, 마할라노비스 거리에 의한 전략의 순으로 나타났다. 이 순서는 상관계수와 root MSE에 대한 척도에서 동일하게 나타났다.

셋째 발견점은 속성수준자료를 그대로 이용하여 거리를 구한 전략 2와 전략 3 보다는 속성의 이상점을 구하여 차원을 축소한 후 거리를 구한 전략 4와 전략 5의 성과가 우수하게 나타났다. 전략 2와

〈표 4〉 실험요인의 상관계수 평균

	삭제 경우 1	삭제 경우 2	삭제 경우 3	평균
전략 1	0.9639	0.9315	0.9002	0.9319
전략 2	0.9651	0.9324	0.8976	0.9317
전략 3	0.9653	0.9268	0.8955	0.9292
전략 4	0.9668	0.9352	0.9078	0.9366
전략 5	0.9696	0.9376	0.9080	0.9384
평균	0.9662	0.9327	0.9018	-

〈표 5〉 실험요인의 root MSE 평균

	삭제 경우 1	삭제 경우 2	삭제 경우 3	평균
전략 1	0.005201	0.008297	0.010558	0.008018
전략 2	0.005329	0.008375	0.010695	0.008133
전략 3	0.005405	0.008796	0.010880	0.008360
전략 4	0.005311	0.008159	0.010238	0.007903
전략 5	0.005069	0.008017	0.010145	0.007744
평균	0.005263	0.008329	0.010503	-

전략 3에서는 공통변수에 속하는 19개의 속성수준을 모두 이용하여 유클리디안 거리나 마할라노비스 거리를 계산한 반면, 전략 4와 전략 5에서는 6개로 축소된 범주형 속성에 대하여 대응일치분석을 적용하여 계량형 좌표를 산출하고 거리를 계산하였다. 전략 2와 전략 3에서는 19차원 상에서 응답자간의 거리를 계산하였는데, 이 때는 분포의 희귀성 때문에 응답자간의 유사성이나 비유사성을 판단하기가 어려워졌기 때문에 예측정확도가 다소 떨어지는 것으로 판단된다.

공통변수의 특징을 잘 포착할 수 있는 방법 중의 하나로 이상점이라는 개념을 도입하여 차원을 6차원으로 축소시켰으며, 이에 의한 방법이 공통변수간의 변별력을 향상시키고 결과적으로 자료융합의 성과가 좋게 나온 것으로 해석할 수 있을 것이다. 즉, 본 연구에서는 정보의 양보다는 정보의 질이 자료융합의 성과에 큰 영향을 미친다고 보인다. 마지막으로 전략 4와 전략 5의 성과 차이는 공통변수에 대한 가정에서 차이점을 찾을 수 있을 것이다. 유클리디안 거리는 변수간의 독립성을 가정하여 계산한 거리인데 반하여, 마할라노비스 거리는 변수간의 종속관계를 나타나는 공분산 구조를 포함하여 응답자간의 거리를 계산하였고 그 결과로 자료융합의 성과를 추가적으로 향상시킬 수 있었다. 마할라노비스 거리가 공통변수간에 실재하고 있는 종속성을 고려함으로써 좀 더 현실을 잘 반영한 방법이라고 볼 수 있을 것이다.

Ⅵ. 결론 및 향후 연구방향

자료융합은 지난 20여 년간 대규모의 데이터 베이스를 다루어야 하는 실무분야에서 많은 주목을 받으며 연구되어왔다. 대규모의 데이터 베이스를 구축함에 있어 개별의 응답자나 고객으로부터 지나치게 많은 양의 정보를 요구하는 것이 현실적으로 불가능하거나 정보의 질을 떨어뜨릴 우려가 있는 경우에 유용한 도구로 사용할 수 있다. 본 연구는 신용카드의 속성에 대한 부분효용자료를 이용하여 자료융합에서 고려할 수 있는 기증자선택전략에 대한 탐색적 연구를 실시하였다. 기존의 방법에서는 상관계수나 유클리디안 거리를 사용하여 기증자를 찾아왔는데, 본 연구에서는 대응일치분석을 적용하였으며, 특히 마할라노비스 거리의 개념을 최초로 자료융합에 적용하였다는데서 의의를 찾을 수 있을 것이다.

자료융합에서 너무 많은 공통변수를 사용하는 것은 자료융합의 성과를 떨어뜨릴 수 있다고 지적된 바 있다(Baker et al., 1997). 또한 인구통계학적 변수나 심리학적 변수를 공통변수로 이용한 것보다 속성자체를 공통변수로 취급한 경우의 자료융합이 더 효과적으로 나타난 연구도 있었다(김성호와 이경미, 1999). 이와 관련하여 본 연구에서는 이상점이라는 개념을 이어받아 공통변수의 차원에 대한 축소를 시도하였고, 유클리디안 거리와 마할라노비스 거리라는 두 가지 측정방법을 도입하였다. 결과는 이전의 연구와 일관성 있는 양상으로 나타났다. 공통변수간의 거리를 직접적으로 계산한 방법보다 공통변수의 차원을 축소하여 거리를 측정하였을 때 자료융합의 성과가 좋게 산출되었다. 또 다른 본 연구의 의의로는 마할라노비스 거리의 개념을 도입한 것이다. 마할라노비스 거리는 기본적으로 유클리디안 거리를 표준화한 것인데, 본 연구에서는 쓰인 부분효용은 각 변수의 scale이 같다는 점을 감안한다면, 표준화의 효과 보다는 변수간의 종속성을 고려한 효과로 인하여 자료융합의 성과가 가장 우수하게 나타난 것으로 판단된다.

자료융합의 주요 적용영역이 시장조사, 선호도 예측, 미디어계획 등임을 감안할 때, 본 연구는 managerial 측면에서는 설문의 설계에 응답자에게 주어지는 설문의 양을 감소시키면서 충실한 답변을 얻어냄으로써 신뢰도 높은 데이터베이스의 구축에 기여할 수 있을 것이다. 또한 설문 및 조사에 소요되는 인력과 시간과 비용 측면에서도 적지 않은 절감을 가져올 것으로 기대된다. 과학적이고 체계적인 자료융합으로 탄생한 고객 데이터베이스는 고객의 니즈파악과 예측 등에 활용되어 direct marketing 등의 다양한 판촉활동의 효과도 높일 것으로 사료된다.

본 연구의 한계점으로는 한 가지의 컨조인트 실험을 통하여 수집한 신용카드 속성에 대한 부분효용 자료에 자료융합을 적용한 것을 들 수 있다. 본 연구의 탐색적 성격을 고려하더라도 연구 결과의 안정성을 높일 수 있도록 다른 실험을 통한 상이한 속성자료에 대한 연구를 시도하여 보아야 할 것이다. 또한 자료융합의 성과를 평가하는 기준도 상관계수나 root MSE 이외의 다양한 기준에 의하여 좀 더 종합적으로 평가할 수 있을 것이다. 각 전략에 대하여도 Scheffe test나 Duncan test와 같은 a posteriori test를 적용한다면 보다 확증적인 연구결론에 도달할 수 있을 것으로 사료된다.

본 연구는 다양한 기증자선택전략을 고안하여 자료융합의 성과를 측정하여 본 탐색적 연구라는 점을 감안할 때, 향후에는 다음과 같은 몇 가지 시도

가 행하여질 수 있을 것이다. 첫째는, 공통변수의 축소에 의한 자료융합의 성과가 본 연구에서 채택한 기증자선택전략에서는 좋게 나타났는데 이러한 현상이 인공지능(artificial intelligence)에 기초한 자료융합의 방법에서도 동일하게 나타나는지 실험해 보는 것은 본 연구가 제시했던 방법론에 대한 탐색적 가치를 증가시킬 수 있을 것이다.

둘째로는, 공통변수의 선택은 각각의 수혜자에 대한 기증자를 지정하는 근거가 되므로 매우 중요하다. 본 연구에서는 12가지의 속성 중에서 분산의 크기가 큰 여섯 가지의 속성을 선택하여 공통변수로 삼았다. 분산의 크기가 큰 속성일수록 기증자를 찾는 과정에서 변별력을 높여줄 수 있으리라는 가정에서 출발하여 이러한 기준을 채택하였다. 향후 연구에서는 다양한 공통변수의 선택기준을 적용하였을 경우에도 본 연구와 동일한 결과를 나타내는지 살펴볼 수 있을 것이다.

셋째로는, 본 연구는 컨조인트 실험에서 얻은 부분효용자료를 사용하였는데 이는 0에서 1사이의 값을 갖는 numeric variable의 형태로 나타났다. 컨조인트 고유의 nonmetric 데이터설계를 사용한다면, 상관계수나 거리의 개념 대신에 association measure를 사용하는 편이 더 적합할 수도 있을 것이다. 즉, 부분효용 이외에 고객의 브랜드 선호도나 만족도, 태도 등의 상이한 성격의 자료를 사용하여 자료융합의 다양한 방법을 적용하고 그 결과를 비교하여 볼 수 있을 것이다.

참고문헌

김성호 · 이경미 (1999), "컨조인트분석에 자료융합방법의 적용에 관한 연구," **마케팅학회**, 14(3), 119-131.

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

Carroll, J. D., P. E. Green, and C. M. Schaffer (1986), "Interpoint Distance Comparisons in Correspondence Analysis," *Journal of Marketing Research*, 23, 271-280.

Carroll, J. D., P. E. Green, and C. M. Schaffer (1987), "Comparing Interpoint Distances in Correspondence Analysis," *Journal of Marketing Research*, 24, 455-470.

Green, P. E. and A. M. Krieger (1993), "Conjoint Analysis with Product Positioning Applications," In J. Eliashberg and G. Lilien (eds.), Handbooks in *OR & MS*, 5, New York: Elsevier Science Publishers.

Green, P. E. and V. R. Rao (1971), "Conjoint Measurement for Quantifying Judgmental Data," *Journal of Marketing Research*, 8, 355-363.

Green, P. E. and V. Srinivasan (1990), "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice," *Journal of Marketing*, 54, 3-19.

Hair, J. F. Jr., R. E. Anderson, R. L. Tatham, and W. C. Black (1998), *Multivariate Data Analysis*, Prentice Hall, New Jersey, 387-437.

Hoffman, D. L. and G. R. Franks (1986), "Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research," *Journal of Marketing Research*, 23, 213-227.

Kamakura, W. A. and M. Wedel (2000), "Factor Analysis and Missing Data," *Journal of Marketing Research*, 37, 490-498.

Kim, J. S., S. Baek, and S. Cho (2004), "A Preliminary Study on Common Variable Selection Strategy in Data Fusion," *Advances in Consumer Research*, 31, 716-720.

Kim, J. S. and M. Hamano (1995), "A Preliminary Study of Data Fusion Approach in Conjoint Analysis," *World Marketing Congress* VII: Proceedings of the Seventh Bi-Annual World Marketing Congress, Melbourne, Australia, 10, 95-101.

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

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

Wittink, D. R. and P. Cattin (1989), "Commercial Use of Conjoint Analysis: An Update," *Journal of Marketing*, 53, 81-86.

Performance Evaluation of Data Fusion Strategy using Mahalanobis Distance

Jonathan S. Kim* · Sungbin Cho**

Abstract

For successful implementation of customer relationship management, corporations today do their best to understand their customer's needs. Conjoint analysis has been used to analyze consumer behaviors toward goods and services. However, it is also notorious for its exponentially increasing number of hypothetical products to evaluate. The most economical and fastest way of gathering information about customers is through questionnaire. Especially today a common form of survey is moving to Internet survey in which respondents can easily get distracted during answering and thereby, survey results might lose sincerity. This study insists that one way of gathering sincere answers from customers is to give them a smaller amount of questions so that they can answer quickly, maintaining constant attention.

Data fusion plays a key role in merging more than two databases and creating an integrated one. Few studies have investigated the intentional data missing where common attribute variables must be determined before collecting data. This study examines various donor location strategies in the area of intentional, preplanned data fusion. We newly introduce the concept of Mahalanobis distance in measuring dissimilarity between respondents in data fusion. In particular, the experiments are accomplished using the following five strategies: Strategy 1 - locating donors by correlation coefficient; Strategy 2 - by Euclidean distance; Strategy 3 - by Mahalanobis distance; Strategy 4 - by Euclidean distance after employing correspondence analysis; and Strategy 5 - by Mahalanobis distance after employing correspondence analysis. By increasing the level of missing, we evaluate the performance of the above five strategies.

* Professor, School of Business, Hanyang University, Seoul, 133-791, Korea
** Assistant Professor, School of Business, Sogang University, Seoul, 121-742, Korea

A part-worth data composed of 12 attributes and 35 attribute levels is used for the experiment. The sample is divided into two groups. The common attribute variables are selected by the size of variance of attributes. The missing variables are randomly selected among the non-common attributes and its all attribute levels are deleted. In the analysis, this type of systematic missing is designed to simulate preplanned data missing. Missing values in one group are substituted from the other group and vice versa.

In the experimental design, the concept of ideal point is introduced. The ideal point means that the maximum attribute level represents the corresponding attribute. Here, continuous variables are converted into categorical variables. To measure the distance from these variables, correspondence analysis is performed in which the coordinates of (p - 1) dimensions are computed where p is the number of attributes. The purpose of ideal point is to decrease the number of dimensions since 19 common attribute levels exist. After applying the correspondence analysis, Euclidean distance and Mahalanobis distance are measured in the last two strategies, compared to purely measuring both distances in Strategies 2 and 3.

A Monte Carlo simulation is conducted 20 times for the five strategies and three levels of missing. The results show that donor location strategy and the level of missing are both statistically significant. By comparing the means of experiment factors, Strategy 5 outperforms other strategies. Next accurate strategy is Strategy 4, and then Strategy 1, Strategy 2, Strategy 3. In the light of the above fact, correspondence analysis seems to play an essential role in decreasing the number of common variables while enhancing explanatory power since Strategies 4 and 5 are better than others.

The purpose of introducing Mahalanobis distance is to reflect the statistical dependence structure of common variables because it has not been considered in the existing Euclidean distance-based data fusion techniques. In the analysis, Mahalanobis distance is effective only when a moderate number of common variables are included in the model. This study explores the possibility of applying Mahalanobis distance for measuring dissimilarity of customers and jointly using correspondence analysis in the area of data fusion. Future extension of this study might apply the proposed donor location strategies into various types of data sets and try to reach a comprehensive conclusion.

Key words: data fusion, Mahalanobis distance, donor location strategy, correspondence analysis