응용통계연구 제 18권 2호, 2005년, pp. 395–410

대규모 분할표 분석*

최 현 집 [1)]

요 약

많은 수의 범주형 변수에 의한 대규모 분할표 분석을 위하여 차원축소(collapsibility) 성질을 이용한 분석 방법을 제안하였다. Kullback-Leibler의 발산 측도(divergence measure)를 이용한 서로 완전한 연관을 갖는 변수그룹을 결정하는 방법을 제안하였으며, 제안된 방법에 의한 변수그룹은 주변 로그선형모형(marginal log-linear models)에 의하여 변수들간의 연관성을 식별할 수 있다. 제안된 방법의 적용 예로 데이터 마이닝에서 흔히 접할 수 있는 대규모 분할표 자료인 소비자들의 구매행위 분석을 위한 장바구니 자료의 분석 결과를 제시하였다.

주요용어: 대규모 분할표, 차원축소, Kullback-Leibler 발산측도, 주변 로그선형모형

1. 서론

Fienberg(2000)가 지적한 바와 같이 컴퓨터 연산능력의 급격한 발전과 범주형 자료분석 이론의 지속적인 발전에 힘입어 연구자들은 대규모의 분할표 분석을 시도하고 있다. 여기서 대규모 분할표(large tables)란 교차분류된 변수의 수는 적으나 각 변수의 범주의 수가 상당히 많거나 혹은 분할표를 구성하는 변수의 수가 많아 분할표가 매우 많은 칸을 갖는 것을 의미한다. 이러한 이유로 대규모 분할표에는 표본크기가 충분히 크더라도 상당부분 칸의 도수가 존재하지 않거나 혹은 매우 작은 도수를 갖는 칸들이 존재할 수 있다.

범주의 수가 매우 큰 둘 혹은 세 변수에 의해 구성된 대규모 분할표에서 의미있는 큰 도수(interestingly large count)를 갖는 칸을 알아내기 위하여 DuMouchel(1999)은 각 칸도수의 절대 크기에 의존하지 않고 비교 가능한 값을 추정하여 관찰칸 값(observed cell counts)과 비교하는 방법을 제안하였다. 이를 위해 칸 도수와 적절한 기대칸 값(expected cell counts)과의 비율인 상대 위험(relative risk), Pearson 통계량으로 부터 얻을 수 있는 표준화 잔차 혹은 각 칸에 대한 우도비 통계량 값을 이용하는 방법 그리고 적절한 사전분포를 고려한 경험적 베이스 측도의 세가지 방법을 제안하였다. 이들 세 방법은 모두 적절한 기대값이 먼저 정의되어야 하며 한 변수가 층을 이루고 나머지 두 변수의 연관만을 고려하여야 하는 경우에 조건부 기대값이 타당한 것으로 제안하고 있다. 또한 셋 이상의 변수에 의한 대규모 분할표에 대해서는 조건부 독립에 의해 설명되는 적절한 그래프 모형(graphical models)에 의한 기대칸 값을 이용할 수 있다고 지적하고 있다. 그러나 많은 수의 변수에 의한 대규모 분

* 이 연구는 2003년 경기대학교 해외파견 연구비 지원에 의하여 연구되었음.

1) (443-760) 경기도 수원시 영통구 이의동 산 94-6, 경기대학교 경제학부 응용정보통계전공, 부교수
E-mail: hjchoi@kyonggi.ac.kr

할표는 고려하여야 하는 모형의 수가 너무도 방대하므로 기대값을 얻기 위한 모형을 선택하는 것은 쉽지 않다. 따라서 많은 변수에 의해 형성된 대규모 분할표로 부터 의미 있는 칸을 식별해내기 위해서는 먼저 적절한 모형을 선택하는 연구가 선행되어야 한다.

같은 맥락에서 Law, Cox, Machonochie, Roman 그리고 Carpenter(2001) 역시 대규모 이차원 분할표를 위하여 경험적 베이스 방법에 의해 의미있는 칸을 식별하는 방법을 제안하였다. 그러나 그들이 제안한 방법은 DuMouchel과는 달리 상대위험들을 정규확률그림에 의해 표현하는 탐색적인 방법을 통해 의미있는 칸을 식별하는데 초점이 맞추어져 있다.

Agresti, Lipsitz와 Lang(1992)은 모두 같은 범주수를 갖는 다수의 범주형 변수에 의한 대규모 분할표의 주변분포를 비교하기 위하여 유사 최우추정량(pseudo maximum likelihood estimator)에 의한 분석방법을 제안하였다. 또한 Erosheva, Fienberg 그리고 Junker(2002)는 같은 상황에서 대규모 분할표 분석을 위한 로그선형모형의 역할과 적용 가능한 여러 다른 모형을 소개하였다. 특히 로그선형모형에 대해서는 준 대칭성(quasi-symmetry)을 이용한 모형이 좋은 분석 도구가 될 수 있으며, 모형에 포함된 모수의 수가 적고 고차 연관항에 포함된 제약조건 때문에 대규모 분할표에서 나타나는 많은 수의 영 칸에 의한 영향을 받지 않을 수 있다고 지적하고 있다. 결국 준 대칭모형(quasi-symmetric models)에 의한 분석은 최우추정에 의해 모형에 포함된 모수를 추정하기 때문에 가능한 저차교호작용항 만을 모형에 포함시켜 차원축소(collapsibility) 성질을 이용하여 축소된 주변표(marginal tables)에 의해 대규모 분할표를 분석하는 방법으로 이해할 수 있다.

대규모 분할표는 데이터 마이닝에서도 흔히 접할 수 있는 자료구조이다. 특히 장바구니 자료분석(market basket analysis)의 경우에 소비자들의 구매행위 분석을 위한 자료는 특정 상품의 구매여부를 이진변수(binary variables)로 한 다차원 분할표로 정리 될 수 있다. 이러한 경우에 Giudici와 Passerone(2002)은 먼저 모든 가능한 두 상품의 구매여부에 따른 이차원 주변표로 부터 승산비(odds ratio)에 따라 연관을 측정하고, 이들 중에서 높은 연관을 갖는 상품들에 의한 변수그룹을 선택하여 분석하는 탐색적인 방법을 제안하였다. 전체 이진변수의 관계는 선택된 변수그룹을 초기모형으로 그래프 모형에 의해 가장 적절한 모형을 선택하는 것으로 연관구조를 파악하게 된다.

이와 같이 다수의 범주형 변수에 의한 대규모 분할표의 분석에 관한 연구의 초점은 다수의 영 칸의 영향을 배제하도록 차원축소 성질을 이용하여 생성된 주변표들을 기반으로 전체 변수들 간의 연관구조를 파악하는데 있다고 판단할 수 있다. 본 연구에서는 이러한 맥락에서 분석에 포함된 전체 변수들 중에서 서로 완전한 연관을 갖는 변수그룹을 결정하고 이들 통해 전체 자료의 연관구조를 식별하는 방법을 제안할 것이다.

제2절에서는 그래프 모형을 기반으로 서로 완전한 연관을 갖는 변수그룹을 결정하기 위하여 Kullback과 Leibler(1951)의 측도를 응용한 유사성 측도에 의한 변수 그룹화 방법을 제안하기로 한다. 또한 제안된 방법에 의한 변수그룹들은 Bergsma와 Rudas(2002)가 제안한 주변 로그선형모형(marginal log-linear models)을 위한 주변분포를 형성하게 하며, 이를 통해 전체 대규모 분할표의 연관구조를 설명할 수 있음을 지적할 것이다. 3절에서는 이들 방법을 적용한 데이터 마이닝에서 흔히 접할 수 있는 대규모 분할표 자료인 소비자들의 구매행위 분석을 위한 장바구니 자료의 분석 예를 제시하였다.

2. 대규모 분할표 분석을 위한 변수 그룹화

적절한 모형에 의한 분할표 분석은 모형에 포함된 모수의 추정값 자체보다는 선택된 모형에 포함된 모수에 의해 분할표를 구성하는 변수들간의 연관관계를 파악하는데 더 많은 관심을 갖게 된다. 이때 조건부 독립성(conditional independence)은 변수들 간의 상호 관련성을 해석하는 좋은 도구가 된다. 그러므로 분석에 포함된 변수를 점(vertex)으로 나타내고 이들 점을 선인 에지(edge)로 연결한 연관 그래프(association graph)를 이용하여 변수들 간의 조건부 독립성을 판단하는 그래프 모형은 분석에 다수의 변수가 포함된 경우에 좋은 분석도구가 될 수 있다.

$V = 5$인 5차원 분할표를 고려하기로 한다. 분할표를 구성하는 변수 A, B, C, D, E에 대하여 이들을 원소로 갖는 집합을 $G = \{A, B, C, D, E\}$로 나타내기로 한다. 만일 $(A, B, C) \perp (D, E)$이라면 즉, 세 변수 A, B, C가 두 변수 D, E와 독립이라면 집합 G에 속한 변수들에 의한 5차원 분할표는 차원축소(collapsibility) 성질에 의해 각각 세 변수와 두 변수에 의한 삼차원 분할표와 이차원 분할표에 의해 설명될 수 있다. 이들 두 변수 집합을 각각 $G_1 = \{A, B, C\}$ 그리고 $G_2 = \{D, E\}$와 같이 나타내기로 한다. 따라서 이들 두 변수 집합에 의하여 전체집합 $G = G_1 \cup G_2$가 될 것이다.

그래프 이론(graph theory)에 의하면 변수를 나타내는 각 점으로 부터 얻을 수 있는 모든 가능한 에지들이 연결되어 있는 경우를 완전한 그래프(complete graph)라 한다. $G_1 \perp G_2$이고 이들 두 집합에 속한 변수들이 각각 모든 가능한 에지들로 연결되어 있다면 두 집합 G_1와 G_2에 속한 변수들에 의한 연관 그래프는 모두 완전한 그래프가 된다. 또한 $G_1 \cap G_2 = \phi$ 이므로 이들 두 그래프는 완전히 분리된 모양을 가질 것이다. 이렇게 전체 집합 G로 부터 서로 분리된 G_1과 G_2를 변수들의 완전한 집합(complete set) 혹은 완전한 변수그룹이라 부르기로 한다.

2.1. 완전히 분리된 변수 집합을 결정하기 위한 측도

집합 G에 속한 변수들의 결합분포함수(joint distribution function)를 p^G 그리고 두 완전한 집합 G_1과 G_2에 의한 결합분포함수를 각각 p^{G_1}과 p^{G_2}로 표현하기로 한다. 즉,

$$\begin{aligned} p^G_{ijklm} &= \Pr(A = i, B = j, C = k, D = l, E = m), \\ p^{G_1}_{ijk} &= \Pr(A = i, B = j, C = k), \\ p^{G_2}_{lm} &= \Pr(D = l, E = m), \end{aligned}$$

여기서 $i = 1, 2, \ldots, \#A$, $j = 1, 2, \ldots, \#B$, $k = 1, 2, \ldots, \#C$, $l = 1, 2, \ldots, \#D$ 그리고 $m = 1, 2, \ldots, \#E$로 각각 변수 A, B, C, D, E의 범주수준을 나타낸다. 이로부터 만일 $G_1 \perp G_2$이라면 전체 5개 변수에 의한 결합분포함수는

$$\begin{aligned} p^{G_1 \perp G_2}_{ijklm} &= p^{G_1}_{ijk} \cdot p^{G_2}_{lm} \\ &= \Pr(A = i, B = j, C = k)\Pr(D = l, E = m) \end{aligned}$$

에 의해 얻을 수 있다. 이때 p^{G_1}과 p^{G_2}는 p^G의 주변분포(marginal distribution)임에 유의하자.

변수집합 G 그리고 G_1, G_2의 상호정보(mutual information)는 G_1과 G_2간의 독립성으로 부터의 거리(distance from independence)로 정의되며 Kullback과 Leibler(1951)의 발산측도(divergence measure)에 의해 측정할 수 있다. 두 분포 p^G와 $p^{G_1 \perp G_2}$에 대한 Kullback과 Leibler의 발산측도는 다음과 같다.

$$\begin{aligned} \mathrm{I}(p^G, p^{G_1 \perp G_2}) &= \sum_{i,j,k,l,m} p^G_{ijklm} \log \frac{p^G_{ijklm}}{p^{G_1 \perp G_2}_{ijklm}} \\ &= \sum_{i,j,k,l,m} p^G_{ijklm} \log(p^G_{ijklm} - p^{G_1 \perp G_2}_{ijklm}) \end{aligned} \tag{2.1}$$

이 정의로 부터 상호정보를 측정하는 발산측도는 대칭(symmetric)이며 항상 양(non negative)의 값을 갖고 만일 $p^G = p^{G_1 \perp G_2}$ 이면 '0'인 것을 알 수 있다. 또한 발산측도는 두 분포 p^G와 $p^{G_1 \perp G_2}$의 거리에 로그를 취한 가중합으로, Kojadinovic(2004)가 지적한 바와 같이 두 분포 p^G와 $p^{G_1 \perp G_2}$의 유사성(similarity)을 측정하는 측도라고 생각 할 수 있다.

집합 G에 속한 변수들이 교차분류된 분할표의 총수를 N 그리고 칸 도수(cell counts)를 n_{ijklm}으로 나타내기로 한다. 분할표 분석에서 결합확률은 칸 비율(cell proportion)을 의미하므로 각 칸의 비율은 $\hat{p}^G_{ijklm} = (n_{ijklm}/N)$에 의해 추정할 수 있다. 그리고 $n_{ijk} = \sum_{lm} n_{ijklm}$과 $n_{lm} = \sum_{ijk} n_{ijklm}$을 각각 변수집합 G_1과 G_2에 의한 주변합(marginal sum)이라고 하면 주변분포 p^{G_1}과 p^{G_2}을 위한 추정 칸 비율은 각각 $\hat{p}^{G_1}_{ijk} = (n_{ijk}/N)$ 그리고 $\hat{p}^{G_2}_{lm} = (n_{lm}/N)$에 의해 추정된다. 그러므로 식 (2.1)은 다음과 같이 추정 할 수 있다.

$$\hat{\mathrm{I}}(p^G, p^{G_1 \perp G_2}) = \sum_{i,j,k,l,m} \hat{p}^G_{ijklm} \log(\hat{p}^G_{ijklm} - \hat{p}^{G_1}_{ijk} \hat{p}^{G_2}_{lm}) \tag{2.2}$$

두 분포의 상호정보를 측정하는 발산측도는 분포 $p^{G_1 \perp G_2}$에 관한 정보를 아는 경우에 분포 p^G의 불확실성의 감소를 측정하는 불확실성 감소측도(uncertainty reduction measure)로 해석될 수 있다. 식 (2.2)의 값이 적절히 크다면 세 변수 A, B, C가 두 변수 D, E와 독립이라는 성질에 의해서 결정된 분포 $p^{G_1 \perp G_2}$에 비하여 전체 변수의 결합분포 p^G의 불확실성 감소량이 크다는 것을 의미하며, 세 변수 A, B, C와 두 변수 D, E의 종속성(dependency)이 존재하는 것으로 해석할 수 있다. 반대로 이 값이 작다는 것은 세 변수가 두 변수와 독립적인 것을 의미하는 것으로 받아 들일 수 있다.

$G_1 \perp G_2$인 경우에 G_1과 G_2는 각각 완전한 그래프를 생성하는 점을 상기하자. 그러므로 이들에 의한 삼차원 주변표와 이차원 주변표를 위해 각각 다음과 같은 로그선형모형을 고려할 수 있다.

$$\begin{aligned} \log p^{G_1}_{ijk} &= \lambda^{ABC} + \lambda^{ABC}_{A(i)} + \lambda^{ABC}_{B(j)} + \lambda^{ABC}_{C(k)} \\ &\quad + \lambda^{ABC}_{AB(ij)} + \lambda^{ABC}_{AC(ik)} + \lambda^{ABC}_{BC(jk)} + \lambda^{ABC}_{ABC(ijk)}, \end{aligned} \tag{2.3}$$

$$\log p^{G_2}_{lm} = \lambda^{DE} + \lambda^{DE}_{D(l)} + \lambda^{DE}_{E(m)} + \lambda^{DE}_{DE(lm)}, \tag{2.4}$$

여기서 λ의 윗 첨자 ABC와 DE는 각각 그룹 G_1과 G_2에 속한 변수에 의한 주변표의 변수를 나타내며, 각 모형에 포함된 λ들은 연관항(interaction term)을 의미한다. 예를 들어 $\lambda^{ABC}_{AC(ik)}$는 G_1에 의한 삼차원 주변표에서의 변수 A와 C의 일차 연관을 나타내는 모수이며 $\lambda^{ABC}_{ABC(ijk)}$는 이차 연관항을 나타낸다. 그리고 $\lambda^{DE}_{DE(lm)}$은 G_2에 의한 이차원 주변표에서의 변수 D와 E의 일차 연관을 나타내는 모수이다. 이들 연관항에는 일반적인 로그선형모형과 같이 모수의 식별을 위한 제약조건, 예를 들어 $\sum_i \lambda^{ABC}_{A(i)} = 0$과 같은 제약조건이 부여된다. 그러므로 두 모형은 각 주변표의 칸의 수와 모형에 포함된 모수의 수가 같은 포화모형(saturated models)이 되는 것을 알 수 있다.

G_1과 G_2는 완전한 집합이고 두 집합에 의한 모형 (2.3)과 (2.4)에 포함되는 모수들은 완전한 계층집합(complete hierarchical sets)을 이루기 때문에 $G_1 \perp G_2$이라는 가정하에서 전체 5차원 분할표의 칸 비율 p^G_{ijklm}은 Bergsma와 Rudas(2002)가 제안한 다음과 같은 주변로그선형모형(marginal log-linear models)으로 나타낼 수 있다.

$$\begin{aligned}\log p^G_{ijklm} &= \lambda^{ABC} + \lambda^{ABC}_{A(i)} + \lambda^{ABC}_{B(j)} + \lambda^{ABC}_{C(k)} \\ &\quad + \lambda^{ABC}_{AB(ij)} + \lambda^{ABC}_{AC(ik)} + \lambda^{ABC}_{BC(jk)} + \lambda^{ABC}_{ABC(ijk)} \\ &\quad + \lambda^{DE}_{D(l)} + \lambda^{DE}_{E(m)} + \lambda^{DE}_{DE(lm)} \end{aligned} \tag{2.5}$$

위 모형은 일반적인 로그선형모형과는 달리 각 주변표에 대한 모형 (2.3)과 (2.4)가 결합된 형태로 다만 모형 (2.4)의 λ^{DE} 만이 제외된 것에 주의 하자. 이러한 이유로 모형 (2.5)에 의한 전체 분할표의 칸 추정값은 $\sum_{i,j,k,l,m} p^G_{ijklm} = 1$ 이라는 조건을 추가한 반복비율적합(IPF; interative proportional fitting) 방법에 의하여 추정하게 된다. 이러한 사실은 만일 $G_1 \perp G_2$이라면 전체 5차원 분할표는 각각 삼차원과 이차원 분할표에 의해 추정된 모수들에 의하여 전체 변수들에 관한 연관구조를 설명할 수 있다는 것을 의미한다.

포화모형에 의한 칸 비율의 추정값은 각 칸의 도수를 분할표의 총 도수로 나눈 값으로 추정할 수 있다. 즉, 집합 G를 위한 포화모형에 의한 칸 비율의 추정값은 $\hat{p}^G_{ijklm} = n_{ijklm}/N$이 된다. 마찬가지로 모형 (2.5)에 의한 칸 추정값은 모형 (2.3)과 (2.4)의 칸 추정값 $\hat{p}^{G_1}_{ijk} = n_{ijk}/N$과 $\hat{p}^{G_2}_{lm} = n_{lm}/N$의 곱으로 추정된다. 그리고 이들은 모두 포화모형의 최우추정값(maximum likelihood estimates)이 된다. 이제 이러한 사실과 식 (2.2)로 부터 다음과 같은 통계량을 고려하기로 한다.

$$\begin{aligned}\mathrm{G}^2(p^G, p^{G_1 \perp G_2}) &= N \sum_{i,j,k,l,m} \hat{p}^G_{ijklm} \log(\hat{p}^G_{ijklm} - \hat{p}^{G_1}_{ijk}\hat{p}^{G_2}_{lm})^2 \\ &= 2N \sum_{i,j,k,l,m} \hat{p}^G_{ijklm} \log \frac{\hat{p}^G_{ijklm}}{\hat{p}^{G_1}_{ijk}\hat{p}^{G_2}_{lm}} \\ &= 2N\hat{\mathrm{I}}(p^G, p^{G_1 \perp G_2}) \end{aligned} \tag{2.6}$$

위 통계량은 모형 (2.5)와 5차원 포화모형 간의 적합도를 평가하기 위한 우도비 검정통계량과 같다. 다시 말해 분할표 분석을 위한 발산측도는 우도비 검정통계량과 관련을 갖는다. 이러한 사실은 분할표의 총 칸도수 N은 항상 영보다 크므로 (2.5) 역시 G_1과 G_2 사이의 독립성으로 부터의 거리를 나타내는 유사성 측도로 이용될 수 있다는 것을 의미한다.

우도비 검정통계량은 모형에 포함된 모수의 수에 영향을 받는다는 것은 널리 알려진 사실이다. 즉, Christensen(1997)이 지적한 바와 같이 일반적으로 로그선형모형을 위한 우도비 검정통계량은 모형에 포함된 모수의 수가 많은면 작아지고 반대로 모수의 수가 적으면 커지는 특징을 갖는다. 이러한 사실은 두 그룹 G_1과 G_2에 포함된 변수의 수와 변수들의 범주의 수에 의해 우도비 검정통계량 값이 영향을 받는다는 것을 의미한다. 따라서 이러한 영향력을 배제하기 위하여 모수의 수에 의해 결정되는 모형의 자유도를 고려한 다음과 같은 측도를 G_1과 G_2 사이의 유사성을 측정하기 위한 측도로 이용하기로 한다.

$$\mathrm{D}(G_1, G_2) = \frac{2N\hat{\mathrm{I}}(p^G, p^{G_1 \perp G_2})}{\mathrm{df}(G_1 \perp G_2)}, \tag{2.7}$$

여기서 우변 분모의 $\mathrm{df}(G_1 \perp G_2)$는 $G_1 \perp G_2$를 검정하기 위한 자유도를 의미하며, 예를 들어 모형 (2.5)를 위한 우도비 검정통계량의 자유도는 $(\#A \cdot \#B \cdot \#C - 1)(\#D \cdot \#E - 1)$이 된다. 그리고 이 값은 $G_1 \perp G_2$인 일반적인 로그선형모형의 자유도와 같다.

2.2. 변수 그룹화를 이용한 대규모 분할표 분석

앞에서 제안한 측도를 이용하여 서로 완전히 분리된 집합을 구성하는 변수집합 혹은 변수 그룹을 결정하는 방법을 설명하기 위하여 Edwards(2000)에서 발췌한 심장질환(coronary heart disease)의 위험요인 자료를 고려해보기로 한다. 자료는 체코슬로바키아에서 1,841명의 운전수를 대상으로 조사되었으며 모두 'yes'와 'no'의 두 범주를 갖는 다음의 6개 변수로 구성되었다.

A : 흡연여부
B : 정신적으로 힘든지의 여부
C : 육체적으로 힘든지의 여부
D : 혈압이 140이하 인지의 여부
E : 혈액내 지단백질의 비율이 3보다 큰지의 여부
F : 직계가족이 관상동맥 심장병을 앓은 경험이 있는지의 여부

만일 이 자료의 6개 변수가 완전 독립성(complete independency)을 만족한다면 서로 완전히 분리된 $\{A\}$, $\{B\}$, $\{C\}$, $\{D\}$, $\{E\}$, $\{F\}$의 6개 완전한 변수그룹이 존재할 것이다. 그러므로 6차원 분할표는 이들에 의해 차원축소된 6개 주변분포에 의해 설명될 수 있을 것이다. 그러나 이들 간에 종속성이 존재한다면, 먼저 이를 식별하기 위하여 Giudichi와 Passerone(2002)이 제안한 것과 같이 6개 변수들로 부터 얻을 수 있는 모든 가능한 2차원 분할표에 의한 부분연관을 만족하는 변수그룹을 설정하는 방법을 고려할 수 있다. 이러한 경우에 6개 변수로부터 얻을 수 있는 모든 가능한 2차원 분할표는 $\binom{6}{2}$개가 존재하며 이들로 부터 계산된 제안된 유사성 측도 (2.7)의 값이 표 2.1에 정리되어 있다.

표 2.1에서 오른쪽 열의 값은 (2.7)에 의해 두 그룹 사이의 유사성을 측정한 한 값이다. 예를 들어 제일 큰 값인 여섯번째 행의 값은 변수그룹 $\{B\}$와 $\{C\}$의 유사성을 측정한 값으

표 2.1: 모든 가능한 두 변수 그룹

변수 그룹	D(·)
$\{A\},\{B\}$	9.664
$\{A\},\{C\}$	27.481
$\{A\},\{D\}$	11.032
$\{A\},\{E\}$	17.400
$\{A\},\{F\}$	1.069
$\{B\},\{C\}$	685.972
$\{B\},\{D\}$	0.500
$\{B\},\{E\}$	17.929
$\{B\},\{F\}$	4.723
$\{C\},\{D\}$	0.095
$\{C\},\{E\}$	16.672
$\{C\},\{F\}$	0.170
$\{D\},\{E\}$	12.809
$\{D\},\{F\}$	1.124
$\{E\},\{F\}$	3.004

로 $\mathrm{D}(\{B\},\{C\}) = 685.972$이다. 이 경우에 각 변수는 모두 두 개의 범주를 가지고 있으므로 $\mathrm{df}(B \perp C) = 1$임에 주의하자.

이 결과로부터 여러 다른 그룹들에 비해 그룹 $\{B\}$와 $\{C\}$의 유사성이 높으므로 즉, 종속성이 강하므로 이들은 동일한 그룹으로 간주하기로 한다. 그러므로 최초 6개 그룹은 $\{B,C\}$, $\{A\}$, $\{D\}$, $\{E\}$, $\{F\}$의 5개 그룹으로 축소 될 수 있다. 이제 이들 5개 변수그룹들로 부터 생성 가능한 그룹의 조합에 대한 표 2.2와 같은 결과를 얻을 수 있다.

앞에서 결정된 완전한 변수그룹 $\{B,C\}$와 이를 제외한 나머지 4개 그룹과의 가능한 각 조합 $\{\{B,C\},A\}$, $\{\{B,C\},D\}$, $\{\{B,C\},E\}$, $\{\{B,C\},F\}$를 위한 자유도는 '3'인 점에 주의하자. 예를 들어 $\{\{B,C\},A\}$에 대하여 $(\#B \cdot \#C - 1)(\#A - 1) = (4-1)(2-1) = 3$이다.

표 2.2의 결과로부터 $\{A\}$와 $\{E\}$의 유사성이 가장 큰 것을 알 수 있다. 그러므로 이들을 새로운 완전한 변수그룹으로 결정하면 원자료를 위한 변수그룹은 $\{B,C\}$, $\{A,E\}$, $\{D\}$, $\{F\}$의 4개가 된다. 같은 방법으로 이들 4개 그룹으로부터 생성 가능한 변수그룹에 대한 유사성을 측정하여 완전한 그룹을 결정하는 방법으로 그룹화를 수행한 결과가 표 2.3에 정리되어 있다.

표 2.3으로부터 이 단계에서 고려할 수 있는 가능한 변수그룹들의 조합 중에서 $\{A,E\}$, $\{D\}$에 대한 $\mathrm{D}(\{A,E\},\{D\})$의 값이 가장 큰 값을 갖는 것을 알 수 있다. 그러므로 이제 $\{B,C\}$, $\{A,D,E\}$, $\{F\}$의 세 완전한 변수그룹을 결정하게 된다. 그리고 역시 이들에 의해 표 2.4의 결과를 얻을 수 있다.

표 2.2: $\{B,C\}$, $\{A\}$, $\{D\}$, $\{E\}$, $\{F\}$로부터의 모든 가능한 변수그룹

변수 그룹	df	D(·)
$\{B,C\},\{A\}$	3	11.157
$\{B,C\},\{D\}$	3	0.443
$\{B,C\},\{E\}$	3	7.700
$\{B,C\},\{F\}$	3	2.340
$\{A\},\{D\}$	1	11.032
$\{A\},\{E\}$	1	17.400
$\{A\},\{F\}$	1	1.069
$\{D\},\{E\}$	1	12.809
$\{D\},\{F\}$	1	1.124
$\{E\},\{F\}$	1	3.004

표 2.3: $\{B,C\}$, $\{A,E\}$, $\{D\}$, $\{F\}$로부터의 모든 가능한 변수그룹

변수 그룹	df	D(·)
$\{B,C\},\{A,E\}$	9	6.879
$\{A,E\},\{D\}$	3	9.783
$\{A,C\},\{F\}$	3	1.597
$\{B,C\},\{D\}$	3	0.443
$\{B,C\},\{F\}$	3	2.340
$\{D\},\{F\}$	1	1.124

표 2.4: $\{B,C\}$, $\{A,D,E\}$, $\{F\}$로부터의 모든 가능한 변수그룹

변수 그룹	df	D(·,·)
$\{B,C\},\{A,D,E\}$	21	3.300
$\{A,D,E\},\{F\}$	7	1.594
$\{B,C\},\{F\}$	3	2.340

표 2.4에서 두 변수그룹 $\{B,C\}$, $\{A,D,E\}$의 유사성이 가장 큰 것을 알 수 있다. 그러나 제안된 방법은 완전한 변수집합을 얻고자 하는 점을 상기하자. 그러므로 변수그룹 $\{A,B,C,D,E\}$가 완전한 변수그룹이 되기 위해서는 5개 변수에 의해 생성 가능한 모든 에지가 포함되어야 한다. 이러한 사실을 확인하기 위하여 표 2.5와 같이 에지들의 포함여부를 결정하기 위한 검정을 수행하기로 한다.

예를 들어 표 2.5의 첫 행의 검정통계량값 $G^2(p^{ABCDE}, p^{A\perp B|CDE}) = 8.74$는 변수 C, D,

표 2.5: 변수그룹 $\{A, B, C, D, E\}$에서의 에지에 대한 검정

에지	$G^2(\cdot,\cdot)$	p-값
$\{A, B\}$	8.74	0.3650
$\{A, C\}$	28.83	0.0003
$\{A, D\}$	20.49	0.0086
$\{A, E\}$	29.72	0.0002
$\{B, C\}$	654.32	0.0000
$\{B, D\}$	5.18	0.7380
$\{B, E\}$	8.91	0.3498
$\{C, D\}$	4.89	0.7690
$\{C, E\}$	10.35	0.2416
$\{D, E\}$	21.91	0.0051

표 2.6: 결정된 변수그룹의 조정된 결정계수

변수 그룹	$R^2(\cdot)$
$\{B, C\}, \{A\}, \{D\}, \{E\}, \{F\}$	0.8095
$\{B, C\}, \{A, E\}, \{D\}, \{F\}$	0.8273
$\{B, C\}, \{A, D, E\}, \{F\}$	0.8555

E가 주어진 변수 A와 B가 조건부 독립인 모형의 적합도 검정통계량값을 의미한다. 이때 각 변수는 두 범주만을 가지고 있으므로 각 자유도는 모두 '1'이 된다. 이 결과로부터 변수 A와 B, B와 D, B와 E, C와 D, C와 E의 에지는 제거될 수 있다는 것을 알 수 있다. 그러므로 변수그룹 $\{A, B, C, D, E\}$은 완전한 변수그룹이 아니다. 따라서 원 6차원 분할표를 위한 완전한 변수그룹은 $\{B, C\}$, $\{A, D, E\}$, $\{F\}$로 결정하게 된다.

이와 같은 과정을 통하여 결정된 각 단계의 변수그룹들이 원 6차원 분할표를 얼마나 설명하는지를 측정하기 위하여 완전 독립성 모형을 최소모형(minimal models)한 Christensen(1997)이 제안한 조정된 결정계수(adjusted coefficient of determination)를 고려하기로 한다. 예컨대 위에서 결정한 세 변수그룹 $\{B, C\}$, $\{A, D, E\}$, $\{F\}$을 위한 다음과 같은 조정된 결정계수를 정의할 수 있다.

$$R^2(\{B, C\}, \{A, D, E\}, \{F\}) = 1 - \frac{\dfrac{G^2(p^{ABCDEF}, p^{BC \perp ADE \perp F})}{\mathrm{df}(BC \perp ADE \perp F)}}{\dfrac{G^2(p^{ABCDEF}, p^{A \perp B \perp C \perp D \perp E \perp F})}{\mathrm{df}(A \perp B \perp C \perp D \perp E \perp F)}} . \qquad (2.8)$$

식 (2.8)을 이용한 앞에서 결정된 각 변수그룹의 조정된 결정계수 값이 표 2.6에 정리되어 있다. 표 2.6으로로부터 세 그룹 $\{B, C\}, \{A, D, E\}, \{F\}$의 조정된 결정계수 $R^2(\{B, C\}, \{A, D,$

$E\},\{F\}) = 0.8555$를 얻게 된다. 즉, 변수그룹 $\{B,C\},\{A,D,E\},\{F\}$에 의해 차원 축소된 주변표가 원 6차원 분할표가 가진 정보를 약 85.5% 설명하는 것으로 해석할 수 있다.

심장질환의 위험요인 자료를 이용하여 얻은 이러한 사실로부터 주어진 대규모 분할표를 구성하는 범주형 변수들을 완전한 집합으로 그룹화하기 위한 다음과 같은 알고리즘을 제안하기로 한다.

1 단계) 주어진 V개 변수로부터 고려할 수 있는 모든 가능한 두 변수를 갖는 그룹을 결정하고 이들 $\binom{V}{2}$개 그룹에 대한 유사성 측도 D(·)을 얻는다. 이 값들 중에서 최대값을 갖는 변수그룹을 결정한다. 이제 결정된 변수그룹에 의해 축소된 이차원 분할표를 이용하여 두 변수의 독립성 검정을 실시한다. 만일 두 변수가 독립인 것으로 판단되면 변수 그룹화를 멈추고 주어진 분할표를 위한 변수그룹은 V개가 존재하는 것으로 판단한다. 그렇지 않은 경우에는 2 단계)를 수행한다.

2 단계) 이전 단계에서 얻은 변수그룹에 속하는 두 변수를 제외한 나머지 변수그룹들과 이전 단계에서 얻은 변수그룹으로부터 생성 가능한 모든 변수그룹을 결정하고, 각 변수그룹에 대한 D(·)을 얻는다. 이 값들 중에서 최대값을 갖는 변수그룹을 결정한다. 새롭게 결정된 변수그룹에서 고려할 수 있는 모든 가능한 각 에지를 제거한 모형에 대한 적합도 검정을 수행하여 만일 이들 적합도 검정 결과 중에서 어떠한 한 검정의 결과에서도 귀무가설을 기각할 수 없다면 이 단계에서 결정된 그룹들을 주어진 자료를 위한 최종 변수그룹으로 판단한다. 그렇지 않으면 다음 단계를 수행한다.

3단계) 그룹화를 위한 반복은 이전 단계와 유사한 방법으로 수행하고, 반복의 종료는 새롭게 결정된 그룹에서 고려할 수 있는 모든 가능한 각 에지를 제거한 모형에 대한 적합도 검정을 수행하여 만일 이들 적합도 검정 결과 중에서 어떠한 한 검정의 결과에서도 귀무가설을 기각할 수 없다면 이 단계에서 결정된 그룹들을 주어진 자료를 위한 최종 완전한 변수그룹으로 판단한다.

알고리즘의 각 단계를 수행하는 과정에서 만일 동일한 유사성 측도 값을 갖는 그룹들이 발견되면, 이들 그룹중에서 모든 가능한 에지를 제거하는 검정을 수행하여 어떠한 한 검정의 결과에서도 에지가 제거되지 않는 그룹을 선택한다. 그러나 만일 모두 에지를 제거할 수 없다면 2.1절에서 지적한 바와 같이 모수의 수가 제안된 유사성 측도의 분자에 영향을 미치므로 자유도가 적은 그룹을 선택하기로 한다.

제안된 완전한 변수그룹을 얻기 위한 알고리즘은 그래프 로그선형모형 중에서 서로 완전히 분리된 완전한 집합을 갖는 모형들만을 고려하는 점에 주의하자. 따라서 최종 선택된 변수그룹들은 원 분할표의 차원보다 낮은 차원의 분할표를 통해 주어진 대규모 분할표를 분석할 수 있게 한다. 그러므로 주어진 대규모 분할표가 많은 수의 영 값을 칸 값으로 갖는 경우에도 결정된 변수그룹에 의한 차원 축소가 이루어지기 때문에 영 값을 갖는 칸들의 영향을 원 분할표에 비해 적게 받을 수 있다.

3. 대규모 분할표의 분석 예

앞에서 제안한 방법으로 SAS/E-Miner의 예제 자료에서 발췌한 고객구매행위 자료를 분석해 보기로 한다. 자료는 아래와 같은 20개 상품에 대하여 1,001명의 고객을 대상으로

표 3.1: 소비자 구매행위 분석의 각 단계에서 선택된 결과

단계	선택된 변수그룹	df	D(·)	R^2
1	$\{H\},\{N\}$	1	356.570	0.0298
2	$\{I\},\{L\}$	1	208.972	0.0473
3	$\{B\},\{C\}$	1	203.774	0.0643
4	$\{J\},\{R\}$	1	177.751	0.0792
5	$\{O\},\{T\}$	1	156.490	0.0923
6	$\{F\},\{M\}$	1	131.595	0.1033
7	$\{D\},\{K\}$	1	85.371	0.1104
8	$\{F,M\},\{Q\}$	3	81.500	0.1308
9	$\{A\},\{S\}$	1	55.004	0.1354
10	$\{I,L\},\{P\}$	3	40.037	0.1455
11	$\{O,T\},\{G\}$	3	36.230	0.1546
12	$\{B,C\},\{D,K\}$	9	33.503	0.1798
13	$\{A,S\},\{I,L,P\}$	21	19.140	0.2133
14	$\{H,N\},\{J,R\}$	9	18.081	0.2269

표 3.2: 변수그룹 $\{G,O,T,B,C,D,K\}$에서의 에지에 대한 검정

에지	$G^2(\cdot,\cdot)$	p-값	에지	$G^2(\cdot,\cdot)$	p-값
$\{B,C\}$	136.64	0.0000	$\{D,G\}$	191.66	0.0000
$\{B,D\}$	57.59	0.0070	$\{D,K\}$	211.67	0.0000
$\{B,G\}$	25.12	0.9129	$\{D,O\}$	241.22	0.0000
$\{B,K\}$	68.72	0.0012	$\{D,T\}$	15.65	0.9970
$\{B,O\}$	54.10	0.0157	$\{G,K\}$	14.66	0.9994
$\{B,T\}$	26.42	0.8200	$\{G,O\}$	290.27	0.0000
$\{C,D\}$	80.61	0.0000	$\{G,T\}$	25.42	0.8827
$\{C,G\}$	42.50	0.1504	$\{K,O\}$	54.21	0.0153
$\{C,K\}$	94.62	0.0000	$\{K,T\}$	100.32	0.0000
$\{C,O\}$	56.35	0.0126	$\{O,T\}$	56.43	0.0123
$\{C,T\}$	35.15	0.4613			

각 고객이 동시에 구입한 상품들을 값으로 가지고 있으며, 각 상품에 대한 구매여부를 이진변수화하면 이들 변수에 의하여 2^{20}차 분할표를 구성할 수 있다.

A : apples	B : artichok	C : avocado	D : baguette	E : bordeaux
F : bourbon	G : chicken	H : coke	I : corned bread	J : cracker
K : ham	L : heineken	M : hering	N: ice cream	O : olives
P : peppers	Q : sardines	R : soda	S : steak	T : turkey

20개 이진변수에 의한 분할표는 모두 1,048,576개 칸을 갖는다. 그러나 이자료는 전체 칸들 중에서 663개 칸 만이 영보다 큰 도수를 갖는다. 그러므로 전체 20개 변수들의 연관관계를 동시에 식별한다는 것은 불가능하다. 이러한 경우에 Giudichi와 Passeron(2002)은 원 분할표를 $\binom{20}{2}$개 이차원 주변표로 차원축소한 이후에 각 이차원 주변표의 승산비(odds ratio)를 통하여 주변연관을 식별하고 이들 중에서 상당히 큰 승산비를 갖는 변수들의 조합으로부터 서로 연관이 있는 변수그룹을 식별하는 탐색적인 방법을 제안하고 있다. 그러나 이 방법에 의해 식별된 변수그룹들은 단지 이차원 주변표의 연관에만 의존하여 그룹화 하였기 때문에 선택된 변수그룹들이 서로 완전한 연관을 가지고 있다고 할 수 없다.

자료가 완전 독립성 모형을 만족한다면 이 자료를 위한 완전한 집합은 서로 완전히 분리된 $\{A\}, \{B\}, \ldots, \{T\}$의 20개 완전한 변수그룹이 존재하는 것으로 생각할 수 있다. 이제 이들 20개 변수그룹을 서로 완전한 연관을 가진 두 변수 이상으로 구성된 변수그룹으로 그룹화하기 위하여 앞절에서 제안한 알고리즘을 적용하고자 한다.

표 3.1으로부터 알고리즘의 첫 단계에서는 $\{H\}$와 $\{N\}$의 유사성이 가장 높은 것으로 결정되었음을 알 수 있다. 그러므로 이들 두 그룹을 완전한 연관이 존재하는 새로운 그룹 $\{H, N\}$으로 결정하게 된다. 반복의 두번째 단계에서는 이 그룹과 나머지 18개 그룹과의 $\binom{19}{2}$개 가능한 변수그룹 중에서 $\{I, L\}$이 새로운 변수그룹으로 선택되었으며, 이러한 과정을 14번 거친 결과가 표 3.1에 정리되어 있다.

제안된 알고리즘을 수행하여 14단계에서 결정된 $\{F, M, Q\}$, $\{G, O, T\}$, $\{B, C, D, K\}$, $\{A, I, L, S, P\}$, $\{H, J, N, R\}$, $\{E\}$의 6개 변수집합으로부터 새로운 반복을 수행하면 $D(\{G, O, T\}, \{B, C, D, K\}) = 8.4111$로 두 그룹 $\{G, O, T\}, \{B, C, D, K\}$의 유사성이 가장 높은 것으로 나타난다. 그러나 이들로부터의 에지 제거를 위한 검정 결과인 표 3.2에서 변수 B와 G의 에지를 포함한 7개의 에지가 제거되어도 되는 것을 알 수 있다. 즉, $\{G, O, T, B, C, D, K\}$은 완전한 집합이 아닌 것으로 판단할 수 있기 때문에 이전단계인 14단계에서 얻은 6개 완전한 집합 혹은 완전한 변수그룹을 최종 그룹으로 결정하게 된다.

각 단계에서 결정된 변수그룹의 전체 자료에 대한 설명 정도를 측정하기 위한 조정된 결정계수 값이 표 3.1의 마지막 열에 제시되어 있다. 이로부터 마지막 단계인 14단계에서 결정된 6개 변수집합에 의해 전체 자료의 약 23% 가 설명되는 것을 알 수 있다. 그러나 주어진 20개 변수에 의한 분할표에는 663개 칸 만이 영 보다 큰 도수를 갖는다는 점을 상기하자. 이러한 이유로 이 자료의 변수들 간의 연관구조를 파악하기 위하여 일반적인 로그선형모형을 적합시키는 것은 현실적으로 불가능하다. 그러나 제안된 방법에 의해 결정된 6개 변수집합들 중에서 가장 많은 변수를 가진 변수집합은 5개 변수로 구성되어 있다. 그러므로 이들 6개 변수집합에 의한 주변 로그선형모형은 많은 수의 영 칸의 영향을 받지 않고 쉽

게 적합할 수 있으며, 이 모형을 통하여 전체 20개 변수들 간의 연관구조를 식별할 수 있게 된다.

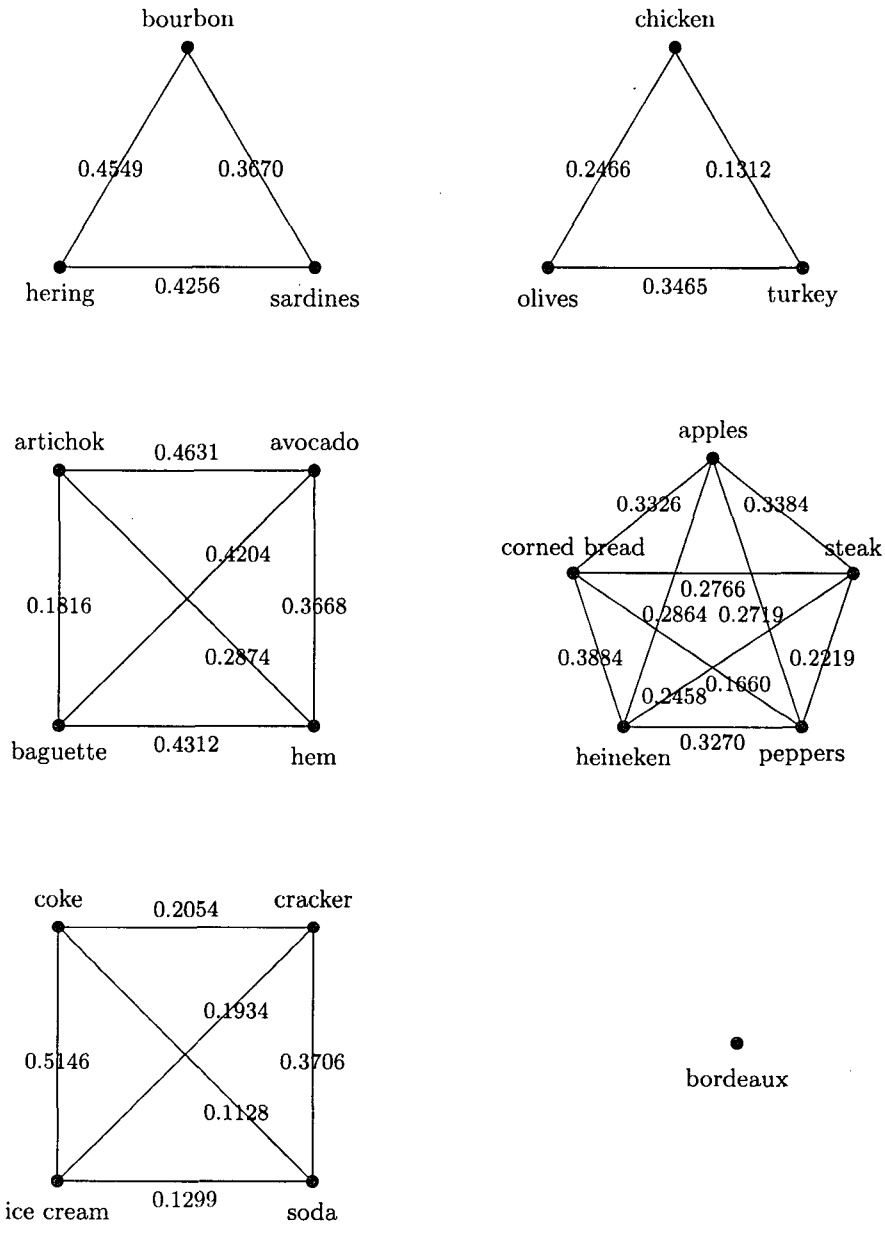

그림 3.1: 선택된 변수그룹들의 연관 그래프

그림 3.1은 이렇게 결정된 그룹들에 의한 주변 로그선형모형의 연관 그래프를 보여주고 있다. 각 연관 그래프의 에지에 나타난 값들은 Whittaker(1990)가 제안한 해당 에지의 강도를 나타내는 부분연관계수(partial correlation coefficients) 값들을 나타낸다. 결과에서 변수 E(bordeaux)만이 어떠 한 상품과도 연관을 가지지 않을 것을 알 수 있으며, 각 변수그룹 즉, 상품그룹들의 연관성이 비교적 고르게 나타난 것을 알 수 있다. 이러한 정보는 Giudici와 Passerone(2002)이 지적한 것과 같이 상품의 프로모션(promotion) 혹은 매장내 상품들의 배치등에 효과적으로 이용될 수 있다.

4. 결론

본 연구에서는 대규모 분할표 분석을 위한 연구들에서 공통적으로 관심을 가지고 있는 차원축소 성질을 이용한 다수의 범주형 변수에 의한 대규모 분할표 분석 방법을 제안하였다.

이를 위하여 먼저, Kullback과 Leibler의 발산 측도를 응용한 유사성 측도를 이용하여 서로 완전한 연관을 갖는 변수들을 선택하는 그래프 모형에 기반한 변수 그룹화 방법을 제안하였다. 제안된 방법에 의해 결정된 변수그룹들은 서로 분리되어 있기 때문에 원 분할표에 비하여 훨씬 낮은 차원에서 변수들 간의 연관을 식별할 수 있게 한다. 따라서 대규모 분할표에 포함된 많은 수의 영 칸의 영향을 배제할 수 있다.

이렇게 결정된 변수그룹들에 의하여 전체 대규모 분할표의 연관구조는 주변 로그선형모형을 통하여 각 변수들간의 관계를 식별할 수 있으며, 기본적으로 그래프 로그선형모형에 기반한 방법이므로 대규모 분할표를 구성하는 변수들의 범주 수에 영향을 받지 않고 적용될 수 있다. 그러므로 준대칭모형을 이용한 분석과는 달리 변수들의 범주수가 같지 않은 경우에도 적용될 수 있다. 또한 준대칭모형에 포함된 연관항들에 주어진 순열(permutation) 성질에 의한 같은 차수의 연관은 모두 같다는 제약을 받지 않으며, 제안된 방법에 의한 변수그룹은 그룹내에 속한 변수들간의 연관구조를 모형에 포함된 모수들에 의해 설명 가능하다는 차이점을 갖는다. 같은 맥락에서 제안된 방법에 의한 변수그룹은 Giudici와 Passerone(2002)이 제안한 방법과 같이 이차원 주변연관을 갖는 것이 아닌 그룹내에 포함된 변수들은 서로 완전한 연관을 갖게 된다.

참고문헌

Agresti, A., Lipsitz, S., and Lang, J. B. (1992). Comparing marginal distributions of large, sparse contingency tables, *Computational Statistics & Data Analysis*, **14**, 55-73.

Bergsma, W. P. and Rudas, T. (2002). Marginal models for categorical data, *Annals of Statistics*, **30**, 140-159.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression 2nd*, Springer-Verlag.

DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system, *The American Statistician*, **53**, 177-190.

Edwards, D. (2000). *Introduction to Graphical Modelling*, Springer-Verlag.

Erosheva, E. A., Fienberg, S. E., and Junker, B. W. (2002). Alternative statistical models and representations for large sparse multi-dimensional contingency tables, *Annales de la Facultè de Sciences de Toulouse*, **11**, 485-505.

Fienberg, S. E. (2000). Contingency tables and log-linear models: Basic results and new developments, *Journal of the American Statistical Association*, **95**, 643-647.

Giudici, P. and Passerone, G. (2002). Data mining of association structures to model consumer behaviour, *Computational Statistics & Data Analysis*, **38**, 533-541.

Kojadinovic, I. (2004). Agglomerative hierarchical clustering of continuous variables based on mutual information, *Computational Statistics & Data Analysis*, **46**, 269-294.

Kullback, S., Leibler, R. A. (1951). On information and sufficiency, *Annals of Mathmatical Statistics*, **22**, 79-86.

Law, G. R., Cox, D. R., Machonochie, N. E. S., E. Roman, J. S., and Carpenter, L. M. (2001). Large Tables, *Biostatistics*, **2**, 163-171.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons.

[2004년 11월 접수, 2005년 4월 채택]

Analysis of Large Tables*

Hyun Jip Choi [1)]

ABSTRACT

For the analysis of large tables formed by many categorical variables, we suggest a method to group the variables into several disjoint groups in which the variables are completely associated within the groups. We use a simple function of Kullback-Leibler divergence as a similarity measure to find the groups. Since the groups are complete hierarchical sets, we can identify the association structure of the large tables by the marginal log-linear models. Examples are introduced to illustrate the suggested method.

Keywords: Large tables, Collapsibility, Kullback-Leibler divergence, Marginal log-linear models

* This research was supported by Kyonggi University Research Grant 2003.

1) Associate Professor, Department of Applied Information Statistics, Kyonggi University, 94-6, Yiui-dong, Yeongtong-gu, Suwon, Kyonggi-do, 443-760, Korea.
E-mail: hjchoi@kyonggi.ac.kr