한국통계학회논문집
2010, 17권, 5호, 667–678

DOI: 10.5351/CKSS.2010.17.5.667

CART 알고리즘을 활용한 확장된 다중인자 차원축소방법의 검정력 평가

이제영[1,a], 이종형[a], 이호근[a]

[a]영남대학교 통계학과

요 약

인간의 유전자 상호작용을 분석하기 위해 제시된 다중인자 차원축소방법은 연속형자료에는 적용할 수 없다. 그래서 이를 보완한 CART 알고리즘을 활용한 확장된 다중인자 차원축소방법이 제안되었다. 하지만 CART 알고리즘을 활용한 확장된 다중인자 차원축소방법의 검정력이 밝혀지지 않았다. 따라서 본 연구에서는 모의실험을 통하여 CART 알고리즘을 활용한 확장된 다중인자 차원축소방법의 우수한 검정력을 평가하고, 확인된 검정력을 바탕으로 실제 한우 데이터에 적용하여 한우의 경제형질에 영향을 주는 우수 유전자조합을 규명하였다.

주요용어: CART알고리즘을 활용한 확장된 다중인자 차원축소방법, 모의실험, 검정력, 한우 경제형질.

1. 서론

인간의 질병 및 가축의 특성 등에 대한 연구에서 단일 유전자의 영향보다 유전자-유전자간의 상호작용에 더 큰 초점을 두고 연구가 진행되고 있다. 특히, 복잡한 특성치들을 연구하는 유전학에서는 유전자의 상호작용에 대한 중요성이 더욱 커지고 있다. 최근 인간 질병에 대한 상호작용을 분석하는 방법으로 Ritchie 등 (2001)과 Chung 등 (2005)이 다중인자 차원축소(Multifactor Dimensionality Reduction; MDR)방법을 제시하였고, 이와 관련된 연구로써 사례-대조로 이분화된 데이터에만 적용 가능한 MDR의 단점을 보완하기 위해서 CART 알고리즘 (Davuluri 등, 2000)을 활용한 확장된 다중인자 차원축소(Expanded MDR; E-MDR, Lee 등, 2008)이 제안되었다. 하지만 E-MDR의 검정력에 대해서는 현재까지 검증된 바 없다.

따라서 본 연구에서는 E-MDR방법의 유전자 상호작용 조합의 선별에 대한 우수한 검정력을 확인하기 위해 모의실험을 통하여 검정력(Power)을 평가하였다. 기존의 MDR방법의 검정력을 확인하기 위해 사용된 연구방법을 바탕으로, 여러 데이터에서 검정력을 평가하기 위해 데이터의 분포와 개체수를 다르게 하여 반복 실험을 하였다. 본 모의 실험에서 사용된 검정력의 평가 측도로는 Bush (2008) 등이 MDR방법의 평가 측도로 사용한 ‘detection’을 사용하였으며, 이는 상호작용을 선별하는 방법이 모든 상호작용의 조합에서 정확하게 선별하는가에 대한 능력을 뜻하는 것으로 본 모의실험에서 정의한 1개의 조합을 얼마나 정확하게 선별(detection)하는지에 대한 정확도(accuracy)를 의미한다. 이러한 ‘detection’에 대항하는 정확도의 평가를 통해서 우리는 E-MDR방법의 검정력을 확인하였다. 아울러

[1] 교신저자: (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 교수. E-mail: jlee@yu.ac.kr

검정력이 확인된 E-MDR의 방법을 실제 한우자료에 적용하여 한우의 등급을 결정하는 경제형질에 관련하는 우수 유전자 조합을 선별하였다.

먼저 2장에서는 상호작용 효과를 규명하는 방법인 MDR방법과 E-MDR방법을 소개한다. 그리고 3장에서는 모의실험을 통해서 E-MDR의 검정력을 평가한다. 4장에서는 모의실험을 통해 검증된 E-MDR의 방법을 실제 한우 자료에 적용하여 경제형질에 영향을 주는 우수 유전자 조합을 선별하고, 마지막 장에서 본연구의 결론 및 연구의 개선점을 제시한다.

2. CART 알고리즘을 활용한 확장된 MDR방법 소개

2.1. MDR방법

MDR방법은 회귀분석과 같은 전통적인 통계모형과는 달리 모수에 대한 추정과 유전형질 모형에 대한 가정이 필요 없다. 즉, 특별한 유전형질 모형(genetic model)에 대한 가정이 필요하지 않다. MDR방법 (Ritchie 등, 2001)은 사례-대조의 이분화된 자료에 적합가능한 방법으로써, 높은 검정력을 가지는 방법이다. 하지만 단지 이분화된 자료에만 적용가능 할 뿐, 연속형 자료에는 적용할 수 없다. 따라서 우리는 연속형 자료에도 적용가능한 방법인 E-MDR방법을 다음 절에서 소개한다.

2.2. CART 알고리즘을 활용한 확장된 MDR방법

E-MDR방법은 이분형 자료에만 적용 가능한 MDR방법을 보완하여 연속형 자료에도 적용하기 위해 제안된 방법으로써 CART방법의 불순도 함수(impurity function)를 사용하여 데이터를 이분화 함으로써 연속형 자료에 MDR방법을 적용할 수 있도록 제시한 기법이다.

아래의 절차와 그림은 연속형 자료에 대한 E-MDR방법의 적용과정을 보여준다.

Step 1. 데이터를 랜덤으로 10개의 같은 크기로 나눈다. 그리고 그 중 9개를 training set으로 1개를 testing set으로 두고, 초기 값을 설정한다.

Step 2. 모든 SNP로부터 k개의 SNP조합 중 하나를 선택한다.

Step 3. 선택된 SNP조합에서 SNP의 각 수준을 기초로 한 개체들을 multifactor classes 또는 cells에 기술한다. 예를 들어서 $k = 2$일 경우, SNP는 3개의 수준으로 되어있으므로 $3^2 = 9$개의 셀을 가진다. 각각 9개의 셀에 경제 형질의 평균값을 기술한다.

Step 4. CART방법의 불순도 함수(impurity function)를 사용하여 연속형 자료를 이분화 한다. 분류 결과로 나온 그룹 중 평균이 높은 그룹이 high가 되고, 평균이 낮은 그룹이 low가 된다.

Step 5. k-factor 조합에 대한 training set에서 잘못 분류된 비율인 평균제곱오차(ASE; average squared error)과, testing set에 대해서 예측된 평균제곱오차(P_ASE; prediction average error)를 아래의 식 (2.1)과 같이 구한다.

Step 6. 모든 SNP의 조합에서 ASE와 P_ASE를 비교하여 가장 작은 조합을 선정한다.

Step 7. Step 1에서 정의한 모든 데이터셋에 대해서 위의 과정을 반복하여 ASE의 평균(Avg.ASE)과 P_ASE의 평균(Avg.P_ASE), CVC(cross validation consistency)값을 구한다. CVC 값은 10번의

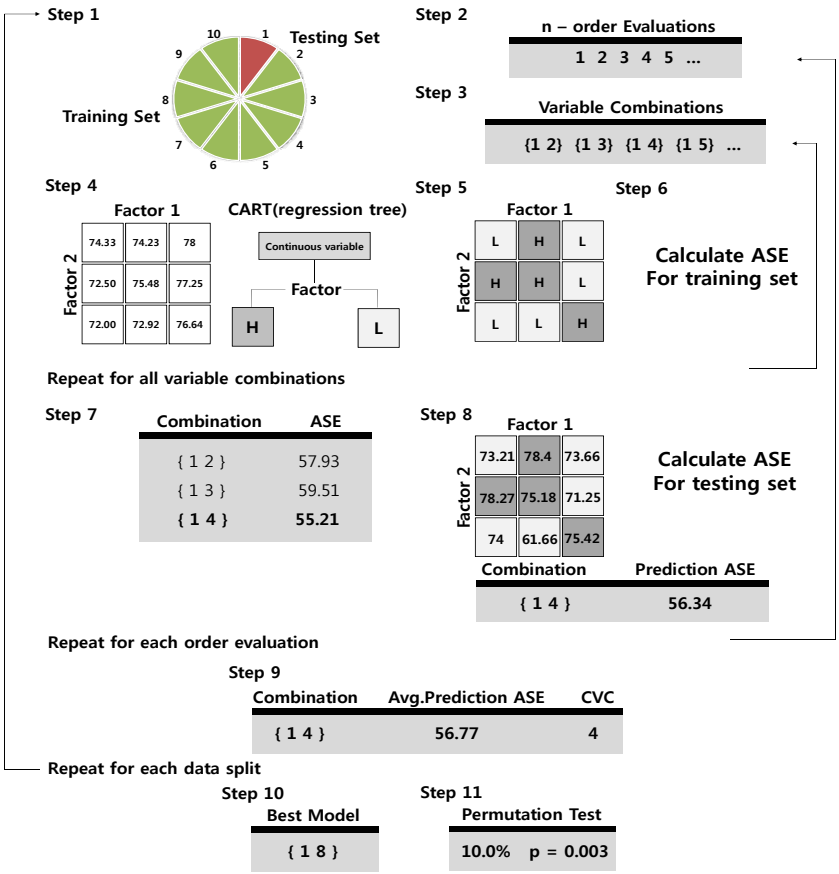

그림 1: *한우의 경제형질에 대한 E-MDR 적용과정*

cross-validation을 시행할 때 각 시행에서 선택된 best model을 카운트 하는 것이다 (Chung 등, 2005).

Step 8. 구해진 Avg.ASE, Avg.P_ASE, CVC 값을 비교해서 Avg.ASE, Avg.P_ASE 값이 낮으며, CVC 값이 큰 조합을 우수 SNP 조합으로 선정한다.

$$\mathrm{ASE} = \frac{S_{high}}{N_{high}} + \frac{S_{low}}{N_{low}}, \tag{2.1}$$

$$S_{high} = \sum_{i=1}^{n} I_{high}(i) \sum_{j=1}^{N_i} (y_{ij} - \hat{y}_i)^2, \qquad S_{low} = \sum_{i=1}^{n} I_{low}(i) \sum_{j=1}^{N_i} (y_{ij} - \hat{y}_i)^2,$$

$$N_{high} = \sum_{i=1}^{n} I_{high}(i) N_i, \qquad N_{low} = \sum_{i=1}^{n} I_{low}(i) N_i,$$

$$I_{high}(i) = \begin{cases} 1, & i(\text{cell}) \in \text{high group}, \\ 0, & \text{o·w}, \end{cases} \qquad I_{low}(i) = \begin{cases} 1, & i(\text{cell}) \in \text{low group}, \\ 0, & \text{o·w}, \end{cases}$$

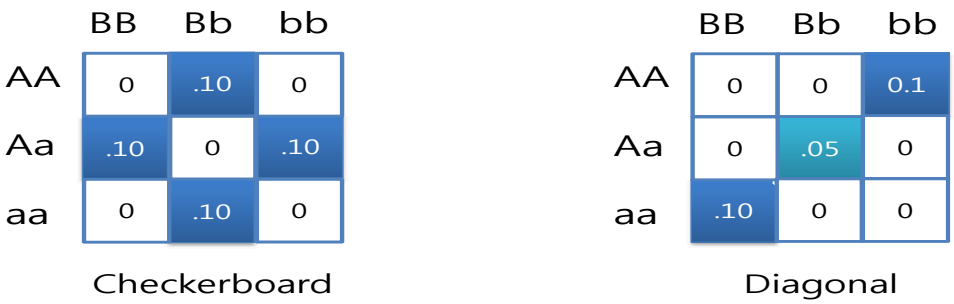

그림 2: *두개의 epistatic 모형(p=0.5, q=0.5)*

표 1: checkerboard모형에서 각 조합과 개별 case의 발현 비율

	각 조합에서 case의 발현 비율			개별 발현 비율
	BB(.25)	BB(.25)	BB(.25)	
AA(.25)	0	1	0	0.5
Aa(.50)	0	1	0	0.5
aa(.25)	0	1	0	0.5
개별 발현 비율	0.5	0.5	0.5	

표 2: diagonal 모형에서 각 조합과 개별 case의 발현 비율

	각 조합에서 case의 발현 비율			개별 발현 비율
	BB(.25)	BB(.25)	BB(.25)	
AA(.25)	0	0	1	0.25
Aa(.50)	0	0.5	0	0.25
aa(.25)	1	0	0	0.25
개별 발현 비율	0.25	0.25	0.25	

3. 모의실험을 통한 E-MDR방법의 정확성 검정

3.1. 모의실험 자료

E-MDR방법의 검정력 평가를 위한 모의실험에서 실제분석자료에 적합한실험을 수행하기 위해 다음과 같은 모의실험 자료를 생성하였다. 모의실험에 활용한 모형은 2가지의 다른 Two-locus epistatic 모형으로써 본 연구와 동일하게 연속형 자료를 분석했던 RPM방법의 모의실험 (Culverhouse 등, 2004)에서 사용된 Checkerboard(바둑판) 모형과 Diagonal(대각) 모형을 활용하여 연속형 데이터 생성을 하였다. epistatic 모형 (Cordell, 2002)을 사용하는 것은 주효과를 제어함으로써 순수한 상호작용의 효과만을 보기 위해서이다. 그림 2와 표 1, 2는 각 모형의 형태와 모형특성을 나타낸다. 표 1과 2에서 조합에 의한 case의 발현 비율은 다르지만 각각 개별적인 형태에서 case의 발현 비율은 동일함을 알 수 있다. 이는 조합에 의한 영향은 존재하지만 개별적인 영향은 없다는 것을 의미한다. 이러한 특성과 Hardy-Weinberg 평형을 유지하는 모형을 epstatic 모형이라 한다 (Ritchie 등, 2003).

본 연구는 연속형 자료에 대한 모의실험으로 각 그룹의 연속형 자료의 분포에 따라 모형의 정확도가 달라질 수 있으므로 총 4가지의 분포에 대해 실험하였다. 각각의 분포는 Culverhouse 등 (2004)이 실험한 RPM의 모의실험에서 사용된 분포 중 4개의 분포를 활용하였으며, 형태는 유지하면서 평균과 표준편차의 경우 실제 적용에 사용될 데이터의 특성을 반영하여 생성하였다. 표 3과 그림 3은 각 분포의 특성과 형태를 나타낸다.

표 3: RPM과 E-MDR 모의실험에서 사용된 분포의 특성

평균	RPM		E-MDR		동일한 영역(%)
	비우수 집단	우수집단	비우수집단	우수집단	
	0	1	312.77	383.93	
SD1	0.25	0.25	17.79	17.79	4.56
SD2	0.50	0.50	35.58	35.58	31.80
SD3	1.00	1.00	71.16	71.16	61.80
SD4	2.00	2.00	142.32	142.32	80.20

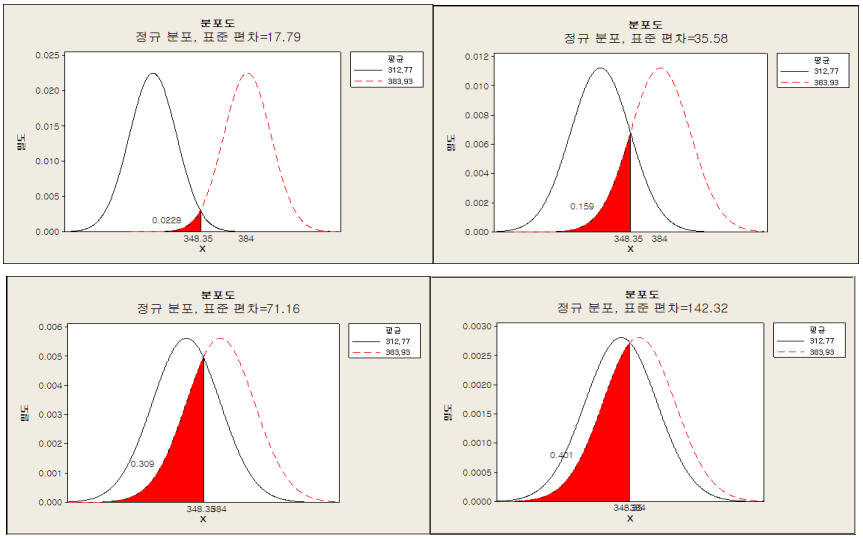

그림 3: *모의실험에서 사용되는 그룹별 각 분포의 형태*

모형의 정확도는 개체수에 따라 차이가 날 수 있으므로 개체수를 3가지의 형태(N = 400, 1000, 2000)로 데이터를 생성하여 실험하였다. 또한 각 실험마다 10개의 SNP를 사용하여 총 45개의 SNP 조합(2개의 조합)을 비교하여 두 모형에 의해 정의된 조합이 얼마나 정확하게 선별되는지를 실험하였다. 각 실험은 10번의 반복을 통해 이루어졌으며, 그룹의 빈도는 균형화되도록 생성되었다. 즉, 각 방법에서 모형(2가지), 표준편차(4가지), 데이터수(3가지), 반복수(10회)를 가지고 총 240개의 데이터 셋(2400개의 cross-validation 데이터 셋)을 생성하여 실험하였다.

그림 4는 SAS ver. 9.13 프로그램을 사용한 데이터 생성 과정을 타나낸다.

Step 1. Epistatic 모형 2가지를 설정하고 초기값을 설정한다.

Step 2. 각 분포에 사용될 평균과 표준편차를 설정한다.

Step 3. 생성될 데이터의 개체수를 설정한다.

Step 4. case와 control의 균형을 맞추기 위해 각 그룹을 $2/N$개의 개체를 가지도록 조건을 설정한다.

Step 5. 실험에 사용될 10개의 SNP(SNP1, . . . , SNP10)에 대한 각 genotype을 생성하고, 2개의 상호작용을 분석하기 위해 45개의 조합 (SNP1 $*$ 2, . . . , SNP9 $*$ 10)에 대해 해당 genotype 조합을 만든다. 이때 각 SNP의 genotype은 가질 수 있는 3개의 형태 중 랜덤으로 선정된다.

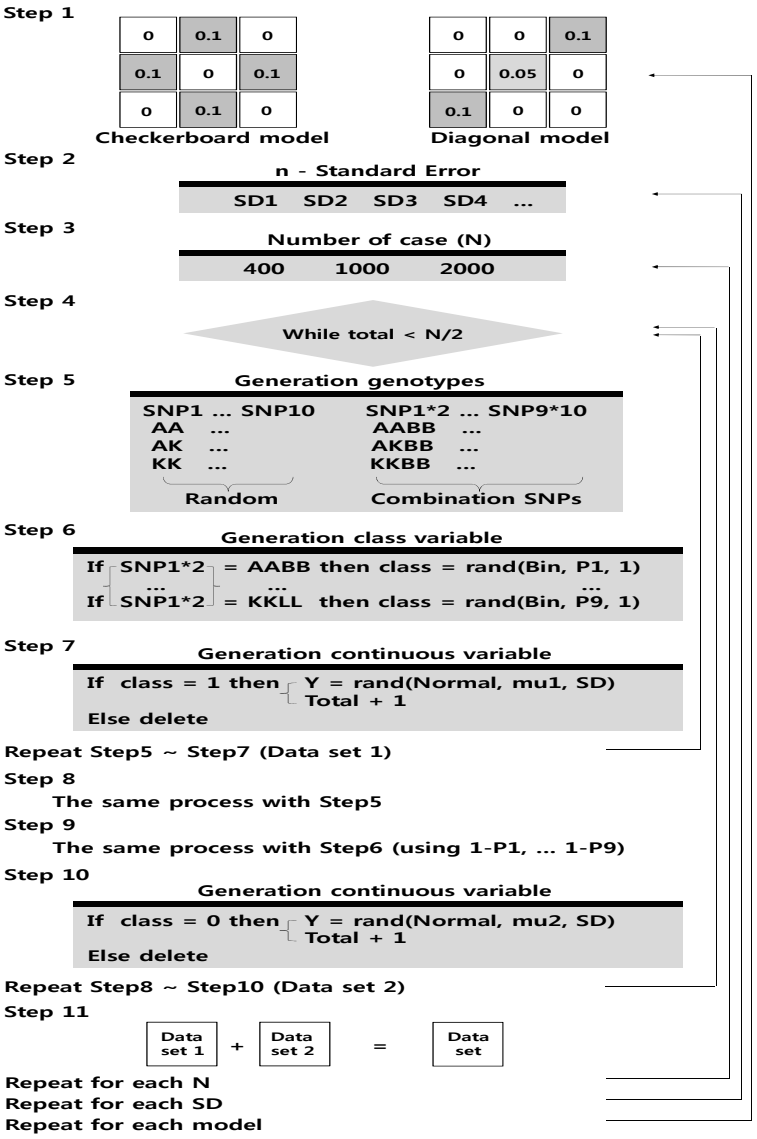

그림 4: *모의실험에서 사용되는 그룹별 각 분포의 형태*

Step 6. 생성된 개체에 그룹에 해당하는 class 변수를 생성한다. class 변수는 오직 SNP1 $*$ 2조합에 의해서 이항분포를 가지고 생성되며, 각 이항분포에서 모수값은 절차1에서 생성한 epstatic 모형에서 설정된 값을 따른다.

Step 7. 상위 그룹에 해당하는 class변수가 1의 값을 가지는 개체에 연속형 자료를 생성한다. 연속형 자료는 절차2에서 설정된 평균(상위그룹에 해당하는 평균)과 표준편차를 가진 정규분포를 따르며, 절차4의 조건을 만족할 때까지 생성하여 데이터 셋(Data set 1)을 생성한다 (두 그룹의 균형을 맞추기 위함).

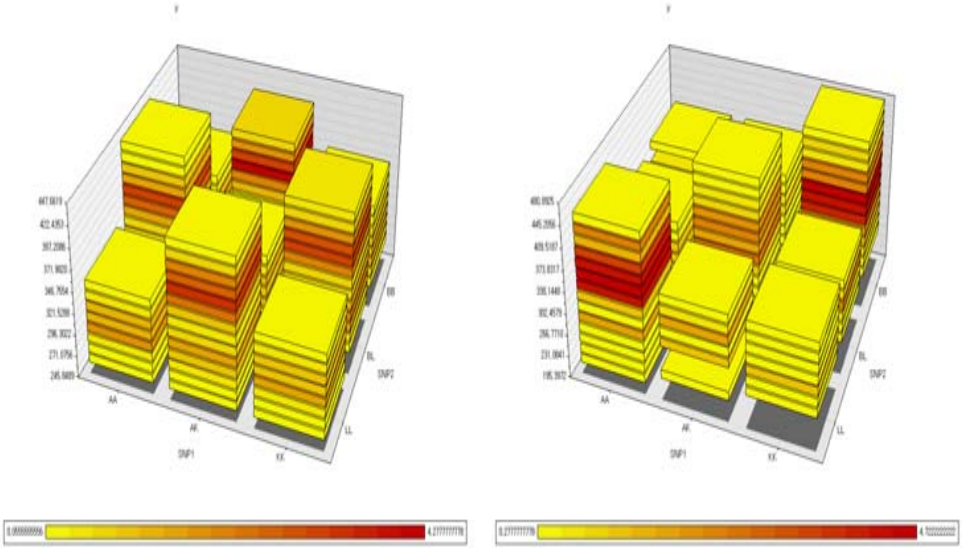

그림 5: *Epistatic 모형으로 생성된 데이터의 분포(n = 2000, SD1)*

Step 8. 하위 그룹 데이터 셋을 생성하기 위한 절차로 절차5와 동일한 수행을 한다.

Step 9. 절차6의 과정을 수행하되, 적용되는 모수의 값($p1, \ldots, p9$)를 하위그룹을 생성하기위한 모수로 변형($1 - p1, \ldots, 1 - p9$)하여 사용한다.

Step 10. class 변수가 0인 개체에 대해 연속형 자료를 생성하는 단계로 하위그룹에 대한 평균을 사용하여 절차7과 같이 수행하여 데이터 셋(Data set 2)을 생성한다.

Step 11. 두 데이터 셋(Data set 1, 2)을 결합하여 데이터셋을 생성하고 절차5에서 절차11의 과정을 절차3, 절차2, 절차1의 순서로 설정을 변경하여 반복한다.

위의 단계를 거쳐 생성된 각 모형의 데이터의 일부 형태는 그림 5와 같이 나타났으며, 생성된 데이터를 모의실험에 사용하였다. 3.2절에서는 모의실험을 통해 알고자 하는 모형의 정확도에 대한 측도를 나타낸다.

3.2. 평가측도

E-MDR방법의 검정력을 평가하기 위한 측도로 Bush 등 (2008) 등이 MDR방법의 평가 측도로 사용한 'detection'을 사용하였으며, 이는 해당 선별 방법이 모든 상호작용 조합에서 정확하게 선별하는 능력을 나타내는 것이다. 본 연구에서는 10개의 SNP의 조합인 45개 조합 중 상호작용 효과로 정의한 1개의 조합을 정확하게 선별하는 정확도(Accuracy)를 말한다. 즉, 각 데이터 셋에서 생성된 그룹과 연속형 자료는 모두 SNP1과 SNP2의 조합인 SNP1 $*$ 2에 의해서 epistatic 모형을 기준으로 생성된 것으로 이 외의 조합은 연속형 자료와 무관하게 무작위 genotype을 가진다. 따라서 E-MDR방법에 각 데이터 셋을 적용하여 분석한 결과 SNP1 $*$ 2가 우수한 상호작용을 가진 조합으로 선택된다면 올바른 선택으로 볼 수 있다. 이러한 평가측도인 정확도를 위해 2개의 epstatic 모형, 4개의 표준편차, 3개의 개체수에 따른 총 24가지 상황별로 10번의 반복에서 정확도를 비교하였다. 3.3절에서는 3.1절에서 생성한 모의실험자료를 E-MDR방법에 적용한 결과를 위에서 나타낸 'detection'에 해당하는 정확도를 통해 평가하고 결과를 통해 나타나는 특성을 나타낸다.

표 4: Checkerboard 모형의 SD3에서 개체수에 따른 반복 결과

표준편차	N	반복	우수 유전자 조합				ACC(%)
			Avg.ASE	Avg.P_ASE	CVC	선택된 우수 유전자 조합	
SD3	400	1	5237.69	5360.35	7	1_2	80
		2	5571.18	5748.87	6	9_10	
		...					
		9	5290.60	5332.00	9	1_2	
		10	5905.45	5965.67	9	1_2	
		평균	5394.95	5533.40	7.30		
	1000	1	6252.91	6317.22	8	1_2	100
		2	6152.39	6287.49	7	1_2	
		...					
		9	5774.54	5854.90	7	1_2	
		10	5640.91	5158.13	7	1_2	
		평균	5951.39	6021.95	7.80		
	2000	1	5785.33	5799.64	10	1_2	100
		2	5863.30	5874.15	10	1_2	
		...					
		9	5919.32	5938.63	9	1_2	
		10	6125.74	6142.18	10	1_2	
		평균	5935.57	5952.88	9.60		

표 5: Checkerboard 모형에 대한 모의실험의 정확도 결과

Checkerboard	개체수	E-MDR Results			ACC(%)
		Avg.ASE	Avg.P_ASE	CVC	
SD1	400	5237.69	5360.35	9.90	100
	1000	1052.76	1056.06	10.00	100
	2000	1074.14	1076.55	10.00	100
SD2	400	2053.85	2072.45	9.50	100
	1000	2006.53	2015.35	9.80	100
	2000	2076.89	2078.71	9.30	100
SD3	400	5394.95	5533.40	7.30	80
	1000	5951.39	6021.98	7.80	100
	2000	5935.57	5952.88	9.60	100
SD4	400	21584.78	22607.43	3.80	20
	1000	21453.99	22607.43	4.80	60
	2000	21094.60	21233.73	4.50	80

3.3. 모의실험 결과

모의실험의 결과는 3.2절에서 설명한 검정력의 측도인 ‘detection’에 해당하는 값을 얻기 위해 E-MDR방법의 절차인 cross-validation 시행을 통해 하나의 결과를 얻고, 그 과정을 10번 반복하여 정확도를 계산하였다.

표 4는 240가지의 실험 중 3가지의 결과에 해당하는 것으로, checkerboard 모형의 표준편차 SD3에서 각 개체수(N = 400, 1000, 2000)에서 10회 반복한 결과이다. 집단에서 10회 반복을 통해 우수 유전자조합이 선택되었으며 정확도는 각각 80%, 100%, 100%로 나타났다. 각각의 경우에서 CVC값을 보면 평균 CVC 값이 7.30, 7.80, 9.60으로 개체수가 많아질수록 정확도와 CVC값이 좋아짐을 알 수 있다.

표 6: Diagonal모형에 대한 모의실험의 정확도 결과

Checkerboard	개체수	E-MDR Results			ACC(%)
		Avg.ASE	Avg.P_ASE	CVC	
SD1	400	965.79	973.96	9.80	100
	1000	920.72	925.18	10.00	100
	2000	938.59	940.84	10.00	100
SD2	400	1916.55	1935.27	9.40	100
	1000	1861.80	1871.23	10.00	100
	2000	1823.27	1827.57	10.00	100
SD3	400	5792.60	5928.11	7.60	100
	1000	5737.93	5804.56	9.40	100
	2000	5733.46	5760.20	9.60	100
SD4	400	20389.55	21166.62	4.20	50
	1000	21182.12	21408.99	5.50	70
	2000	20881.25	21001.19	5.40	80

표 5과 6에서 각 epistatic 모형에서 모의실험 결과를 종합하여 나타낸다. 표 5와 같이 모의실험 결과 모형의 정확도를 확인하였다. 두 epistatic 모형에서의 결과는 큰 차이는 없었으나 Diagonal 모형에서의 정확도가 조금 더 높았으며, 두 그룹의 표준편차가 작을수록 CVC값과 정확도가 증가하였다. 또한 개체수가 증가 할수록 정확도 역시 증가하였다.

다음장에서는 모의실험결과를 토대로 개체수의 보완을 위해 Bootstrap (Efron과 Tibshirani, 1993) 법을 사용해 개체수를 보완하여 우수한 상호작용 효과를 지닌 유전자 조합을 선별한다.

4. E-MDR방법 적용 결과

4.1. 실험 자료

본 연구의 데이터는 농협중앙회 한우개량사업소의 후대검정집단인 30차에서 35차 국가 후대검정우 집단 476두로 구성된다. 한우의 경제형질인 일당증체량(ADG)과 도체중량(CWT), 등심단면적(LMA), 근내지방도(MS)에 영향을 주는 6개의 htSNP(haplotype-tagging SNP)를 (Lee, 2009) 사용하여 E-MDR방법을 적용하였다.

4.2. 적용 결과

한우의 경제형질에 대한 E-MDR적용 결과 SNP의 단일효과에서 g.34425+102A>T가 공통적으로 선정되었다. 또한 2개의 SNP조합의 경우 표 8과 같이 g.8778G>A, g.11500-117C>G의 조합이 일당증체량과 근내지방도에서 우수 조합의 선정되었고, 등심단면적과 도체중량에서는 각각 g.8778G>A, g.66995-169insdelC 조합과 g.11500-117C>G, g.34425+102A>T 조합이 선정되었다. 표 9에서는 3개의 SNP조합에 대한 결과를 나타내었다. 일당증체량과 근내지방도에서 2개 SNP조합과 같이 g.8778G>A, g.11500-117C>G, g.34425+102A>T이 선정되었고, g.8778G>A, g.32330-48A>G, g.34425+102A>T 조합과 g.11500-117C>G, g.34425+102A>T, g.66995-169insdelC 조합이 각각 등심단면적과 도체중량에 영향을 주는 우수 유전자 조합으로 선정되었다. 또한 E-MDR방법의 적용결과 선정된 유전자형 조합의 통계적 유의성에 대해서 검정하기 위해 순열검정을 실시하였다 (Good, 2000). 그 결과 개별 SNP효과에서는 근내지방도를 제외한 나머지 형질에서 통계적인 유의성을 입증할 수 없었다. 이는 개별적인 SNP의 효과보다 이들의 상호작용에 의한 영향이 더 크다는 것을 말해준다. 또한 E-MDR의 방법이 단일효과보다는 상호작용의 효과를 규명하는데 유용하다는 것을 짐작하게 한다.

표 7: 개별 SNP에 대한 E-MDR 적용 결과

경제형질	우수 유전자 조합	ASE	P_ASE	순열검정(*P*-value)
일당증체량(ADG)	g.34425+102A>T	0.007958	0.007970	0.10100
근내지방도(MS)	g.34425+102A>T	16.3377	16.3539	0.01411
등심단면적(LMA)	g.34425+102A>T	58.5674	58.6149	0.13481
도체중량(CWT)	g.34425+102A>T	1146.223	1146.866	0.07347

표 8: 2개 SNP조합에 대한 E-MDR 적용 결과

경제형질	우수 유전자 조합	ASE	P_ASE	순열검정(*P*-value)
일당증체량(ADG)	g.8778G>A g.11500-117C>G	0.007796	0.007696	0.00194
근내지방도(ADG)	g.8778G>A g.11500-117C>G	16.1373	16.3072	0.00424
등심단면적(ADG)	g.8778G>A g.66995-169insdelC	58.2487	57.0311	0.00327
도체중량(ADG)	g.11500-117C>G g.34425+102A>T	1120.013	1118.707	0.03898

표 9: 3개 SNP조합에 대한 E-MDR 적용 결과

경제형질	우수 유전자 조합	ASE	P_ASE	순열검정(*P*-value)
일당증체량(ADG)	g.8778G>A g.11500-117C>G g.34425+102A>T	0.007755	0.007621	0.00737
근내지방도(ADG)	g.8778G>A g.11500-117C>G g.34425+102A>T	15.9860	16.1578	0.00367
등심단면적(ADG)	g.8778G>AT g.32330-48A>G g.34425+102A>T	57.6772	56.7029	0.01640
도체중량(ADG)	g.11500-117C>G g.34425+102A>T g.66995-169insdelC	1111.296	1109.692	0.00406

5. 결론 및 토의

광범위 유전자 관련(genome-wide association; GWA)연구에서 수많은 유전자들을 이용하여 인간의 질병에 관련된 유전자를 찾는 방법 MDR방법이 제시되었다. 우리는 본 연구에서 이분형 자료에만 적용 가능한 MDR방법의 한계를 극복하기 위한 방법으로 제시된 E-MDR방법을 소개하고, E-MDR방법에 의한 우수 유전자 조합 선별결과에 대해 신뢰성 등을 확인하기 위해 모의실험을 통해 검정력을 평가하였다. 정확도(accuracy)에 해당하는 선별능력(detection)을 평가하였으며, 모의실험 결과 평균 정확도 89.1%로 높게 나타났다. 특히 개체수가 충분히 큰 경우($n > 1000$), 93.13%, 그룹의 분포에서 겹치는 면적이 약 61.0%미만 (SD3 이내) 일 경우 99.16%의 높은 정확도를 확인 할 수 있었다. 즉, 모의실험 결과 우수 상호작용 조합을 선별하는 방법 E-MDR방법의 검정력을 확인하였다.

또한 모의실험을 통해 검정된 E-MDR방법을 적용하여, Lee 등 (2009)이 haplotype 분석을 통해 선별한 한우의 경제형질에 우수한 6개 SNP로 한우의 등급을 결정하는 경제형질에 해당하는 일당증체량과 도체중, 등심단면적, 근내지방도에 영향을 주는 우수한 SNP 조합을 선별하였다. 그 결

과 한우의 경제형질에 관련된 우수 유전자 조합중 2개의 조합에서 일당증체량와 근내지방도의 경우 g.8778G>A, g.11500-117C>G, 등심단면적과 도체중량에서 각각 g.8778G>A, g.66995-169insdelC의 조합과 g.11500-117C>G, g.34425+102A>T 조합이 선정되었고, 3개의 SNP조합에서는 일당증체량와 근내지방도의 경우 g.8778G>A, g.11500-117C>G, g.34425+102A>T의 조합, 등심단면적은 g.8778G>A, g.32330-48A>G, g.34425+102A>T, 도체중량은 g.11500-117C>G, g.34425+102A>T, g.66995-169insdelC 조합이 각각 우수한 유전자 조합으로 선정되었다. 따라서 E-MDR에 의해 선정된 우수 유전자 조합을 잘 활용한다면 육량과 육질에서 높은 등급을 가진 한우를 생산 할 수 있을 것으로 기대한다.

참고 문헌

Bush, W. S., Edwards, T. L., Duek, S. M., McKinney, B. A. and Ritchie, M. D. (2008). Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction, *Bio Medical Central Bioinformatics*, **9**, 238.

Chung, Y. J., Lee, S. Y. and Park, T. S. (2005). Multifactor dimensionality reduction in the presence of missing observations, *Proceedings of the Autumn Conference Korea Statistical Society*, 31–36.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans, *Human Molecular Genetics*, **11**, 2463–2468.

Culverhouse, R., Klein, T. and Shannon W. (2004). Detecting epistatic interactions contributing to quantitative traits, *Genetic Epidemiology*, **27**, 141–152.

Davuluri, R., Suzuki, Y., Sugano, S. and Zhang, M. (2000). CART Classification of Human 5′ UTR Sequences, *Genome Res*, **10**, 1807–1816 .

Efron, B. and Tibshirani, R. (1993). *An introduction to the Bootstrap*, Springer-Verlag Berlin and Heidelberg GmbH & Co, New York.

Good, P. (2000). *Permutation Test: A ractical Guide to Resampling Method for Testing Hypotheses*, Chapman & Hall/CRC, New York.

Lee, J. Y., Kwon, J .C. and Kim, J. J. (2008) Multifactor dimensionality reduction(MDR) analysis to detect single nucleotide polymorphisms associated with a carcass trait in a Hanwoo population, *Asian - Australasian Journal of Animal Sciences*, **6**, 784–788.

Lee, Y. S. (2009). Study on the identification of candidate genes and their haplotypes that are associated with growth and carcass traits in the QTL region of BTA6 in a Hanwoo population, *Ph. D. Thesis, Yeungnam University*, 1–94

Ritchie, M. D., Hahn, L. W. and Moore, J. H. (2003). Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phencopy, and Genetic Heterogeneity, *Genetic Epidemiology*, **24**, 150–157.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl F. F. and Moore, J. H. (2001). Multifactor dimensionality reduction reveals high-order interactions among estrogen- metabolism genes in sporadic breast cancer, *American Journal of Human Genetics*, **69**, 138–147.

2010년 6월 접수; 2010년 7월 채택

Power of Expanded Multifactor Dimensionality Reduction with CART Algorithm

Jea-Young Lee[1,a], Jong-Hyeong Lee[a], Ho-Guen Lee[a]

[a]Department of Statistics, Yeungnam University

Abstract

It is important to detect the gene-gene interaction in GWAS(Genome-Wide Association Study). There are many studies about detecting gene-gene interaction. The one is Multifactor dimensionality reduction method. But MDR method is not applied continuous data and expanded multifactor dimensionality reduction(E-MDR) method is suggested. The goal of this study is to evaluate the power of E-MDR for identifying gene-gene interaction by simulation. Also we applied the method on the identify interaction effects of single nucleotide polymorphisms(SNPs) responsible for economic traits in a Korean cattle population (real data).

Keywords: CART algorithm, E-MDR, economic traits, power.

[1] Corresponding author: Professor, Department of Statistics, Yeungnam University, Kyungsan 712-749, Korea.
E-mail: jlee@yu.ac.kr