한국통계학회논문집
2009, 16권, 4호, 579-586

연속형 데이터에서 E-MDR과 D-MDR방법 비교

이제영[1,a], 이호근[a]

[a]영남대학교 통계학과

요약

통계모형의 상호작용 효과를 분석하기위해 비모수적인 방법인 다중인자 차원 축소(MDR)방법을 사용해왔다. MDR 방법은 사례-대조 데이터에만 적용 할 수 있다. 본 논문에서는 Regression tree 알고리즘과 더미변수를 활용한 회귀분석 알고리즘을 사용하여 다중 범주를 High범주와 Low범주로 분류함으로써, MDR방법에서 연속형 데이터에 적용 할 수 없는 문제를 해결하는 방법으로 제시된 Expanded MDR방법과 Dummy MDR방법을 한우의 주요 경제형질(longissimus muscle dorsi area: LMA, carcass cold weight: CWT, average daily gain: ADG)데이터에 적용하여 한우의 경제형질에 영향을 주는 주요 SNPs 마커를 규명하고, Permutation test를 통해 그 결과를 비교한다.

주요용어: Dummy MDR, expanded MDR, SNP, 한우 경제형질.

1. 서론

통계모형의 상호작용을 고려한 모형으로 선형모형 같은 표준 통계적 모형을 사용해왔다. 그러나 유전자와 같은 범주형 데이터의 경우 변수가 많을 경우 상호작용의 조합이 많아지므로 종종 모수들의 상호작용에 대한 해석과 모형을 결정하는 것이 어려울 수 있다. 그래서 다중인자 차원 축소(Multifactor Dimensionality Reduction: MDR)방법 (Ritchie 등, 2001; Chung 등, 2005), 조합 분할(combinatorial partition method: CPM)방법 (Nelson 등, 2001), 제한된 분할(restricted partition method: RPM)방법 (Culverhouse 등, 2004) 등이 여러 유전자에 대한 상호작용을 결정하는 방법들로 개발되었다. 특히 MDR방법은 상호작용에 대한 명확한 모형의 가정이 없는 비모수적인 방법으로 적당한 high-order 차수의 데이터로 복잡한 관계를 밝힐 수 있었다. MDR방법은 사례항목과 대조항목의 비율을 통해 독립변수의 다중 범주를 'high'범주와 'low'범주로 분류한 후 목표변수에 대한 오분류율을 비교하여 분석한다. 그러나 이 MDR방법은 사례-대조로 이분화 된 데이터에 대해 사용하는 방법으로 연속형 데이터에 대해 적용하는데 문제점이 발생한다. 이 문제점을 연속형 목표변수로부터 독립변수의 다중 범주를 'high'범주와 'low'범주로 분류하여 해결하기 위해, CART(classification and regression tree)알고리즘을 활용하는 방법인 Expanded MDR 방법 (Lee 등, 2008)과 더미변수를 활용한 회귀분석 알고리즘을 활용하는 방법인 Dummy MDR 방법 (Lee와 Kim, 2009)이 제시되었다. 본 논문에서는 이 두가지 방법을 한우의 경제형질 데이터에 적용하여 한우의 경제형질에 영향을 주는 주요 SNPs 마커를 규명하고, Permutation test (Good, 2000)를 통해 그 결과를 비교한다.

본 연구는 2009년도 영남대학교 학술연구 조성비에 의한 것임.

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

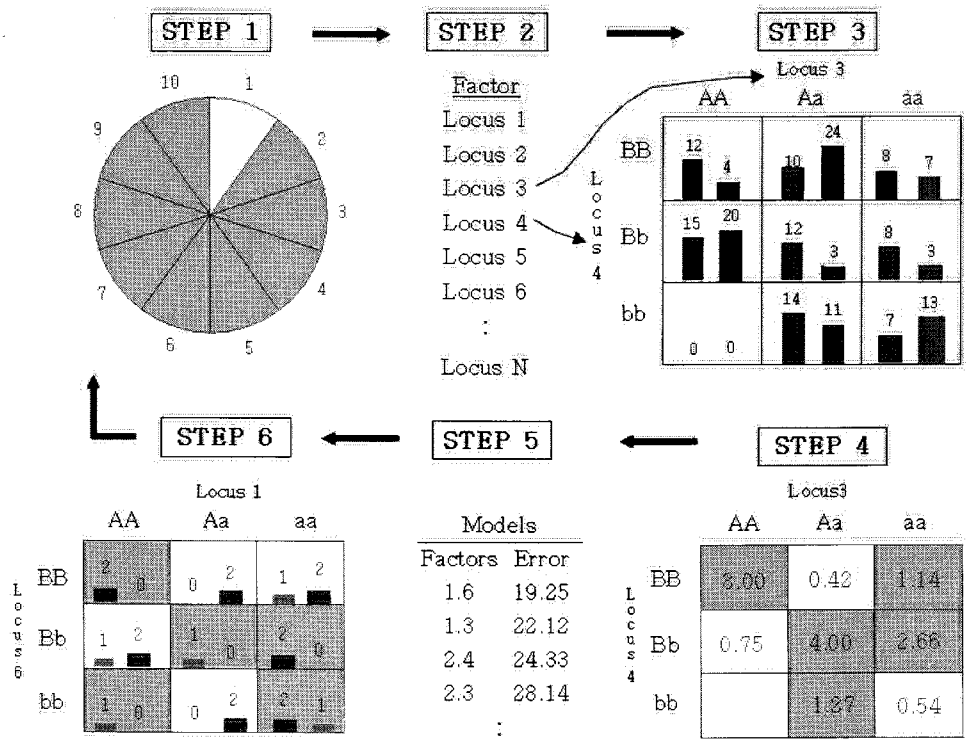

그림 1: *case-control 데이터에서 MDR방법의 적용과정*

2. E-MDR과 D-MDR방법

사례-대조 데이터에만 적용 가능한 기존의 MDR방법을 소개하고, 연속형 데이터에 적용하기 위해 제시된 Expanded MDR (Lee 등, 2008), Dummy MDR방법 (Lee 등, 2009)의 특징을 소개한다. 아울러 두 분석의 결과를 비교하기 위해 사용한 Permutation test방법을 소개한다.

2.1. MDR방법

MDR 방법은 일반화된 선형 모형인 전통적인 통계 기법과는 달리 어떤 모수에 대한 추정과 genetic 모형의 가정을 요구하지 않는다(다시 말하면 특별한 유전형질 모형에 대한 가정이 필요 없다). Ritchie 등 (2003)과 Hahn 등 (2003)에 의하면 MDR 모형에서 처음으로 시행하는 것은 case와 control을 1:1로 균형을 맞추는 것이다. 그림 1과 다음 step에서 case-control에 대한 MDR 방법을 시행하는 과정을 보여준다.

Step 1. 데이터를 랜덤으로 10개의 같은 크기로 나눈다. 그리고 그 중 9개를 training set으로 1개를 testing set으로 둔다.

Step 2. 모든 SNP로부터 k개의 SNP조합 중 하나를 선택한다.

Step 3. 선택된 SNP조합에서 SNP의 각각 수준을 기초로 한 개체들을 multifactor classes 또는 cells에 기술한다. 예를 들어서 $k = 2$일 경우 SNP는 3개의 수준으로 되어 있다. 따라서 $3^2 = 9$개의 셀을 가진다. 각각 9개의 셀에 case의 값과 control의 값을 적는다.

Step 4. Case와 control의 비를 구하여 1보다 크거나 같으면 high-risk, 1보다 작으면 low-risk라 한다. 예를 들면 1행 1열의 경우 case와 control의 비가 1보다 작으면 이 셀은 low-risk이다.

Step 5. K개의 SNP의 조합 전부에서 데이터의 9/10인 training set에서 잘못 분류된 비율인 misclassification error(ME)를 구한다. 여기서 잘못 분류된 비율인 ME는 $\{(\text{Total}_{high} - \text{Case}_{high}) + \text{Case}_{low}\}/N$이다. Total_{high}는 high그룹의 전체 값이며 Case_{high}는 high그룹의 전체 case 경우의 수이며 Case_{low}는 low그룹의 전체 case 경우의 수이다. 그리고 N은 전체 데이터의 수이다. 이렇게 구한 ME들 중에 가장 작은 값을 선택한다.

Step 6. Training set에서 high-low로 나눈 표를 나머지 1/10의 데이터인 testing set을 이용하여 잘못 분류된 비율인 prediction error(PE)를 Step 5의 정의와 같이 구한다.

그 다음 위의 과정의 반복에서 나온 10개의 ME와 PE의 평균을 구해 그 값이 가장 낮은 것을 best n-factors 모형으로 정한다 (Bastone 등, 2004). 그리고 앞에서 구한 각각의 ME를 이용하여 cross validation consistency(CVC)를 구하는데 이것은 10번의 cross- validation을 시행할 때 각 시행에서 선택된 best model을 카운트하는 것이다 (Chung 등, 2005). 따라서 ME와 PE의 평균이 가장 낮고 CVC가 가장 높은 값이 best n-factors 모형이다.

2.2. E-MDR 방법

Expanded multifactor dimensionality reduction(E-MDR)방법은 case-control 데이터에만 적용 가능한 MDR방법의 문제점을 해결하기 위해 제안된 방법으로 CART(classification and regression tree)알고리즘을 이용한 MDR방법의 확장이다 (Lee 등, 2008). 2.1절에서 소개한 MDR과정은 절차에서 Step 3과 4의 과정에서 다중범주를 high집단과 low집단으로 이분화하기 위해 목표변수가 case-control 형태가 되어야 한다. 따라서 목표변수가 연속형 데이터인 경우 이분화가 불가능하게 된다. E-MDR방법에서는 Step 3과 4의 절차를 CART알고리즘을 통한 이분화 방법을 적용함으로서 연속형 자료인 경우에도 MDR방법을 적용 할 수 있도록 제시한 기법이다.

2.3. D-MDR 방법

목표변수가 연속형일 경우 MDR방법에 적용 할 수 없는 문제를 해결하기 위해 제시된 다른 방법으로 더미(dummy)변수를 활용한 Dummy with multifactor dimensionality reduction(D-MDR)방법 (Lee 등, 2009)이 있다. D-MDR방법은 문제를 해결하기 위해 더미변수를 활용한 회귀분석방법 (김태근, 2006)을 활용한다. 즉, 다중범주를 더미변수로 변환하여 회귀분석을 시행하여, 더미변수의 회귀계수들의 상호관계를 통해 이분화 하는 알고리즘을 D-MDR방법과 같이 2.2절의 MDR방법 절차에서 Step 3과 4의 절차에 이 알고리즘을 적용하는 방법이다. E-MDR방법과 D-MDR방법 모두 평가방법으로 2.1절에서 소개된 misclassification error(ME)와 prediction error(PE)를 대신하여 average squared error(ASE)와 prediction average squared error(P-ASE)를 사용한다 (Lee 등, 2008; Lee 등, 2009).

2.4. Permutation test 방법

두 방법의 결과를 비교하기 위해 본 논문에서는 Permutation test (Good, 2000)를 통한 p-value를 사용한다. 다음의 절차에 따라 Permutation test를 시행하였다.

절차 1. 검정하고자하는 가설 설정 - 2.3절의 절차를 통해 선정된 우수 SNP 마커가 경제 형질에 영향력이 있다.

절차 2. 통계량과 기각역 설정 - 분석에 사용할 통계량으로 'high'그룹의 평균을 선택. 선정된 SNP 마커에 의한 'high'그룹의 평균이 그룹간 데이터를 서로 바꾸었을 때 보다 높다면 경제형질에 영향력이 있다고 판단한다.

절차 3. 기존 관측치의 통계량 계산 - 각 경제형질 데이터(검정용 데이터셋)를 더미변수를 활용한 MDR모형에 의해 선택된 'high'그룹과 'low'그룹으로 이분화 시킨 후 'high'그룹의 평균을 계산한다.

절차 4. 관측치의 재배열과 재배열 후의 통계량 계산 - 두 그룹의 데이터를 n개 만큼 랜덤 추출하여 그룹을 상호 변경한 후 각 'high'그룹의 평균을 구한다. 이 과정을 10,000번 반복한다.

절차 5. 결론 - 각 평균을 내림차순으로 정렬한 후 기존의 평균과 비교하여 Monte Carlo의 유의확률 값을 구한다. 절차 2-4과정을 10번 시행(10개의 훈련용 데이터셋)하여 유의확률들의 평균값을 구하여 유의수준보다 작으면 선정된 우수 SNP 마커가 경제 형질에 영향력이 있다는 가설을 채택한다.

위의 순열 검정을 통해 E-MDR방법과 D-MDR방법에 의해 선정된 SNP 마커들에 대한 p-value를 비교하게 된다. 3장에서는 2장에서 살펴본 방법들을 적용한 결과를 나타낸다.

3. E-MDR방법과 D-MDR방법의 적용 및 결과

3.1. 실험자료

본 연구 데이터는 농협중앙회 가축개량 사무소에서 개발되었고 16grand-sire half-sibs families로부터 229두의 수송아지로 구성되어졌다. 한우의 여러 경제형질인 등심단면적(LMA: longissimus muscle dorsi area), 도체중(CWT: carcass cold weight), 일당증체량(ADG: average daily gain)은 모든 F1 자손으로부터 수집되어졌고 한국축산물등급판정소의 규격에 따라 측정되었다.

현재까지 소에서는 도체형질(도체중량, 등지방두께, 등심단면적, 일당증체량, 근내지방도)과 연관이 있는 SNP 마커(marker)들이 일반가축에서 평가되어지거나 적용되고 있다 (Barendse 등, 2004; Page 등, 2004). 따라서 본 연구에서는 EST-based SNP 연관지도 (Snelling 등, 2005)에서 Kim 등 (2003)에 의해 규명 되어진 한우 염색체 6번에 위치한 후보 QTL인 ILSTS035와 같은 거리에 있는 SNP들 중 다형성이 나타난 SNP(19_1, 18_4, 28_2)를 이용하였다.

3.2. E-MDR방법의 적용 결과

등심단면적(LMA), 도체중(CWT), 일당증체량(ADG)에 대해 SNP의 조합에 대하여 E-MDR 과정을 10번 반복해서 나온 ASE와 P-ASE의 평균과 10번의 반복과정에서 나온 값을 기준으로 CVC를 구한 결과는 각각 표 1, 2와 같이 나타냈다. 표 1, 2를 통해 모든 경제형질에서 하나의 SNP에 의한 효과보다 두 개의 SNP조합에 의한 효과가 더 좋은 것으로 나타났으며, SNP(19_1)과 SNP(28_2)의 조합에 의한 효과가 가장 좋은 것으로 나타났다.

3.3. D-MDR방법의 적용 결과

등심단면적(LMA), 도체중(CWT), 일당증체량(ADG)에 대해 SNP의 조합에 대하여 D-MDR 과정을 10번 반복해서 나온 ASE와 P-ASE의 평균과 10번의 반복과정에서 나온 값을 기준으로 CVC를 구한 결과는 각각 표 3, 4와 같이 나타났으며, E-MDR방법과 마찬가지로 하나의 SNP에 의한 효과보다 두

표 1: E-MDR방법에 의한 average ASE와 average P-ASE 결과

요인의수	SNP 마커	LMA		CWT		ADG	
		ASE	P_ASE	ASE	P_ASE	ASE	P_ASE
1	**SNP** (19_1)	**58.872**	**59.251**	**1074.21**	**1087.07**	**.007262**	**.007218**
	SNP(18_4)	61.268	61.479	1100.73	1111.79	.007410	.007384
	SNP(28_2)	60.615	60.875	1093.52	1099.03	.007335	.007275
2	SNP(19_1)*SNP(18_4)	58.514	58.311	1071.66	1076.53	.007215	.007253
	SNP (19_1)*SNP (28_2)	**57.875**	**57.409**	**1056.55**	**1044.73**	**.007003**	**.006990**
	SNP(18_4)*SNP(28_2)	59.862	59.160	1069.90	1080.04	.007203	.007313

표 2: E-MDR방법에 의한 CVC 결과

요인의수	SNP 마커	LMA	CWT	ADG
1	**SNP** (19_1)	**10**	**10**	**10**
	SNP(18_4)	0	0	0
	SNP(28_2)	0	0	0
2	SNP(19_1)*SNP(18_4)	0	0	0
	SNP (19_1)*SNP (28_2)	**10**	**10**	**10**
	SNP(18_4)*SNP(28_2)	0	0	0

표 3: D-MDR방법에 의한 average ASE와 average P-ASE 결과

요인의수	SNP 마커	LMA		CWT		ADG	
		ASE	P_ASE	ASE	P_ASE	ASE	P_ASE
1	**SNP** (19_1)	**59.051**	**58.501**	**1089.17**	**1085.70**	**.007311**	**.007269**
	SNP(18_4)	60.548	60.039	1102.76	1095.87	.007465	.007427
	SNP(28_2)	60.652	60.139	1103.75	1100.68	.007463	.007423
2	SNP(19_1)*SNP(18_4)	58.351	58.043	1080.12	1070.78	.007243	.007242
	SNP (19_1)*SNP (28_2)	**57.793**	**57.429**	**1043.67**	**1031.85**	**.006990**	**.006947**
	SNP(18_4)*SNP(28_2)	59.110	59.239	1077.68	1075.50	.007302	.007263

표 4: D-MDR방법에 의한 CVC 결과

요인의수	SNP 마커	LMA	CWT	ADG
1	**SNP** (19_1)	**10**	**10**	**10**
	SNP(18_4)	0	0	0
	SNP(28_2)	0	0	0
2	SNP(19_1)*SNP(18_4)	0	0	0
	SNP (19_1)*SNP (28_2)	**10**	**10**	**10**
	SNP(18_4)*SNP(28_2)	0	0	0

SNP의 조합에 의한 효과가 더 좋은 것으로 나타났으며, SNP(19_1)과 SNP(28_2)의 조합에 의한 효과가 가장 좋은 것으로 나타났다.

3.2절과 3.3절의 결과를 보면 주요 경제형질에 대해 영향이 가장 많은 SNPs 조합 마커는 E-MDR과 D-MDR에 의한 결과가 동일하게 나타났으며, 영향력이 가장 좋은 SNPs 마커로 나타난 SNP(19_1)* SNP(28_2) 조합 마커의 average ASE와 average P-ASE 값을 보면 D-MDR에 의한 값이 좀 더 낮은 것을 확인 할 수 있다. 3.4절에서는 E-MDR과 D-MDR방법에 의해 나누어진 결과를 Permutation test의 결과를 통해 비교한다.

표 5: E-MDR방법에 의한 Permutation test 결과

SNP 마커	순열 검정 결과 유의확률 값		
	등심단면적	도체중	일당증체량
SNP(19_1)	0.04900	0.04264	0.03651
SNP(19_1)*SNP(28_2)	0.00658	0.00996	0.00484

표 6: D-MDR방법에 의한 Permutation test 결과

SNP 마커	순열 검정 결과 유의확률 값		
	등심단면적	도체중	일당증체량
SNP(19_1)	0.03506	0.04574	0.03579
SNP(19_1)*SNP(28_2)	0.00214	0.01125	0.00255

3.4. 순열 검정 결과

2.3절에서 소개한 Permutation test를 하나의 SNP에 의한 효과 중 가장 높은 것으로 나타난 SNP (19_1)과 두 SNP 조합에 의한 효과 중 가장 높은 것으로 나타난 SNP(19_1)*SNP(28_2)에 대해 적용한 결과를 표 5와 표 6에 나타내었다. Permutation test의 결과를 보면 유의수준 0.05에서 모두 유의한 것으로 나타났으며, 하나의 SNP에 의한 효과보다 두 SNP의 조합에 의한 효과가 더 유의하게 나타났다. E-MDR방법과 D-MDR방법 모두 동일한 결론을 나타내고 있음을 알 수 있으며, 두 방법의 p-value값을 보면 등심단면적(LMA)과 일당증체량(ADG)에서 D-MDR에 의한 결과의 p-value가 좀 더 낮은 것을 확인 할 수 있다.

4. 결론 및 토의

이분형 데이터에만 적용 가능한 MDR방법의 문제점을 해결하기 위한 방법인 E-MDR방법과 D-MDR방법을 소개하고 이 방법들을 이용하여 연속형 데이터로 이루어진 한우의 경제형질 데이터에 적용 하였다. E-MDR방법과 D-MDR방법에 적용하여 ASE, P_ASE, CVC값을 통해 분석한 결과 두 방법 모두 동일한 결론인, 단일 SNP marker에서는 SNP(19_1) marker가 한우의 경제형질에 영향을 많이 주었으며, 두 개의 SNP조합 marker에서는 SNP(19_1)*SNP(28_2) marker가 한우의 경제형질에 영향을 가장 많이 주는 것으로 나타났으며 SNP(19_1) marker와 SNP(19_1)*SNP(28_2) marker의 ASE와 P-ASE를 비교한 결과 SNP(19_1) marker인 하나의 유전자의 효과보다는 SNP(19_1)*SNP(28_2) marker와 같이 조합에 의한 유전효과가 한우의 경제형질에 더 많은 영향을 주는 인자로 나타났다. 두 방법의 결과에 대해 비교한 결과로는 D-MDR방법에 의한 ASE와 P-ASE의 값이 좀 더 낮은 것으로 나타났으며, Permutation test 결과에서도 등심단면적과 일당증체량에서 좀 더 유의한 것으로 나타났다.

참고 문헌

김태근 (2006). <u-Can 회귀분석>, 인간과 복지, 서울.

Barendse, W., Bunch, R., Thomas, M., Armitage, S., Baud, S. and Donaldson, N. (2004). The TG5 thyroglobulin gene test for a marbling quantitative trait loci evaluated in feedlot cattle, *Australian Journal of Experimental Agriculture*, **44**, 669–674.

Bastone, L., Reilly, M., Rader, D. J. and Foulkes, A. S. (2004). MDR and PRP: A comparison of methods for high-order genotype-phenotype associations, *Human Heredity*, **58**, 82–92.

Chung, Y. J., Lee, S. Y. and Park, T. S. (2005). Multifactor dimensionality reduction in the presence of missing observations, *Journal of Korea Statistical Society, Proceedings of the Autumn Conference*, **1**,

31–36.

Culverhouse, R., Tsvika, K. and William, S. (2004). Detecting epistatic interactions contributing to quantitative traits, *Genetic Epidemiology*, **27**, 141–152.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman & Hall/CRC.

Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, Springer-Verlag Berlin and Heidelberg GmbH & Co., New York.

Kim, J. W, Park, S. I. and Yeo, J. S. (2003). Linkage mapping and QTL on chromosome 6 in Hanwoo(Korean Cattle), *Asian-Australasian Journal of Animal Sciences*, **16**, 1402–1405.

Lee, Y. S., Bae, J. H., Lee, J. Y., Park, H. S. and Yeo, J. S. (2008). Identification of candidate SNP for economic traits on chromosome 6 in Korean cattle, *Asian-Australasian Journal of Animal Sciences*, **21**, 1703–1709.

Lee, J. Y. and Kim, D. C. (2009). Important SNPs identification from the economic traits for the high quality Korean cattle, *Communications of the Korea Statistical Society*, **16**, 67–74.

Nelson, M. R., Kardia, S. L. R., Ferrell, R. E. and Sing, C. F. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation, *Genome Research*, **11**, 458–470.

Page, B. T., Casas, E., Quaas, R. L., Thallman, R. M., Wheeler, T. L., Shackelford, S. D., Koohmaraie, M., White, S. N., Bennett, G. L., Keele, J. W., Dikeman, M. E. and Smith, T. P. L. (2004). Association of markers in the bovine CAPNI gene with meat tenderness in large crossbred populations that sample influential industry sires, *Journal of Animal Science*, **82**, 3474–3481.

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.

Snelling, W. M., Casas, E., Stone, R. T., Keele, J. W., Harhay, G. P., Benett, G. L. and Smith, T. P. L. (2005). Linkage mapping bovine EST-based SNP, *BMC Genomics*, **6**, 74–84.

2009년 4월 접수; 2009년 6월 채택

A Study on the Comparison between E-MDR and D-MDR in Continuous Data

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

[a]Department of Statistics, Yeungnam University

Abstract

We have used multifactor dimensionality reduction(MDR) method to study interaction effect of statistical model in general. But MDR method cannot be applied in all cases. It can be applied to the only case-control data. So, two methods are suggested E-MDR and D-MDR method using regression tree algorithm and dummy variables. We applied the methods on the identify interaction effects of single nucleotide polymorphisms(SNPs) responsible for longissimus mulcle dorsi area(LMA), carcass cold weight(CWT) and average daily gain(ADG) in a Hanwoo beef cattle population. Finally, we compare the results using permutation test.

Keywords: Dummy MDR, expanded MDR, Hanwoo economic traits, single nucleotide polymorphism.

This research was supported by the Yeungnam University research grants in 2009.

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