한국통계학회논문집
2011, 18권, 4호, 413–422

DOI: 10.5351/CKSS.2011.18.4.413

연속형 자료의 유전자 상호작용 규명을 위한 SVM MDR과 D-MDR의 방법 비교

이종형[a], 이제영[1,a]

[a]영남대학교 통계학과

요 약

유전학에서 유전자 상호작용효과 규명을 위한 방법으로 비모수적인 방법인 Multifactor Dimensionality Reduction(MDR) 방법이 제안되어 현재까지 사용되고 있다. MDR 방법은 이분형 자료에 적합한 방법으로 연속형 자료에는 적용할 수 없는 단점이 있다. 이러한 한계를 극복하기 위해서 Dummy MDR(D-MDR) 방법 그리고 SVM을 활용한 MDR(SVM MDR) 방법 등이 제안 되었다. 본 논문에서는 연속형 자료에 적용 가능한 SVM MDR 방법과 D-MDR 방법을 비교하고, 실제 한우 데이터에 두 방법에 적용한다. 그리고 각 방법의 적용결과를 바탕으로 한우의 종합경제형질에 영향을 주는 유전자 상호작용 조합을 규명한다. 그리고 마지막으로 기존의 SVM MDR 방법과 D-MDR 방법의 장단점 비교를 통해서 추후 새로운 연구방향을 제시한다.

주요용어: Dummy MDR, Multifactor Dimensionality Reduction(MDR), SVM, 종합경제형질.

1. 서론

유전학에서는 유전자 단일 효과보다 상호작용에 의한 효과의 해석이 더 필요하다. 이는 단순한 표준 통계적 모형으로는 해석하기에 한계가 있다. 일반적으로 변수의 증가는 모형의 복잡도를 증가시키고 복잡도가 증가한 모형은 해석이 매우 어려워진다. 또한 모수적인 방법에 의한 모형을 사용함에 있어 모형에 대한 가정을 만족하지 못하는 경우가 많다. 이러한 여러 문제점들을 해결하기 위해서 유전자 상호작용을 규명하는 방법으로 비모수적인 방법인 다중인자 차원축소 방법(Multifactor Dimensionality Reduction; MDR)이 제안되었다 (Ritchie 등, 2001).

비모수적인 방법인 MDR 방법은 모형에 대한 어떠한 가정도 필요치 않으며, 고차원의 데이터를 적당한 차원으로 축소함으로써 변수들 사이의 관계를 쉽게 밝힐 수 있다. 최근 MDR 방법은 인간의 질병 및 가축의 특성에 대한 연구에서 많이 사용되고 있으며, 복잡한 특성치를 규명하는 유전학분야에서 좋은 성능을 발휘한다. 또한 MDR 방법은 이분형으로 분류된 자료에 적용이 가능한 방법으로 높은 검정력을 평가 받았다. 하지만 MDR 방법은 연속형 자료에는 적용할 수 없다. 단지 이분형 자료에만 적용이 가능할 뿐이다. 이와 같은 MDR 방법의 한계를 보완하기 위한 연구들이 많이 진행되어 왔으며, 현재까지도 진행되고 있다. Dummy 변수를 활용한 다중인자 차원축소 방법 (Dummy MDR; D-MDR; Lee 등, 2009a)이 대표적인 그것이다.

본 연구에서는 연속형 자료에 적용 가능한 MDR 방법으로 이분형 분류기술인 Support Vector Machine(SVM) 알고리즘을 이용한 MDR 방법과 D-MDR 방법의 비교를 통해서 각 방법의 장점과 단점을 제시한다. 아울러 한우 데이터에 위의 두 방법을 적용하여 얻어진 결론을 바탕으로 한우의 경제형질과 연관된 단일염기다형성(SNP)들의 상호작용을 규명한다. 그리고 선별된 유전자 상호작용의 조합에 대해서 통계적 유의성을 확인한다. 먼저 2장에서는 먼저 유전자 상호작용 규명을 위한 방법인 MDR 방

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

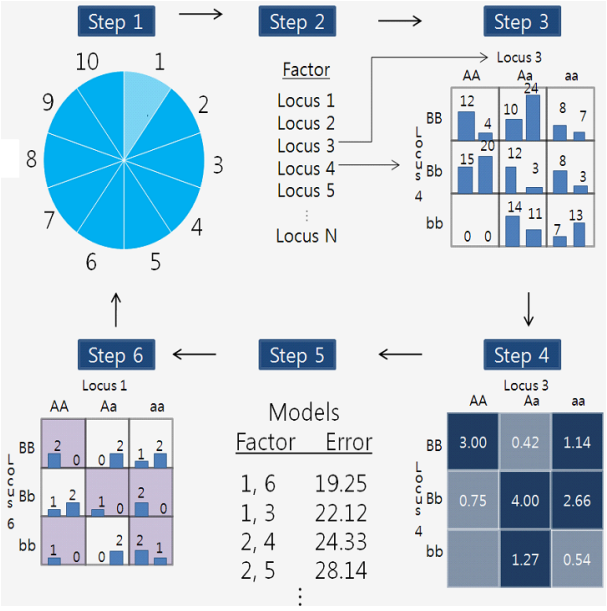

그림 1: *Case-Control 데이터에서 MDR 방법의 적용과정*

법에 대해서 소개한다. 그리고 이분형 분류기술인 SVM 알고리즘과 SVM을 활용한 MDR 방법, 더미 변수를 활용한 MDR 방법인 D-MDR 방법에 대해서 소개한다. 3장에서는 실제 한우 데이터에 SVM을 활용한 MDR 방법의 결과와 D-MDR 방법을 적용한 결과를 비교하여 한우의 경제형질에 연관된 우수 유전자 상호작용 조합을 선별하고, 선별된 우수 유전자 조합에 대해서 통계적 유의성을 확인하기 위해 순열검정을 적용한다. 마지막으로 4장에서는 본 연구의 결론을 제시하여 각 방법의 장점과 단점을 통해 향후 연구에 대해 논의한다.

2. SVM MDR과 D-MDR의 방법비교

2.1. MDR 방법

MDR 방법은 일반화된 선형 모형의 전통적 통계기법과는 달리 모수에 대한 추정과 모형에 대한 가정이 필요하지 않은 비모수적 방법이다. 또한 MDR 방법은 검정력 평가를 통해서 높은 검정력을 보인 우수한 유전자 상호작용 효과 규명을 위한 방법이다. MDR 방법에서 가장 선행되어야 하는 것은 바로 case와 control의 비를 1 : 1로 맞추는 것이다. 다음의 Step과 그림 1은 case-control에 대한 MDR 방법의 시행과정을 보여준다.

Step 1. 데이터를 크기가 동일한 10개의 셋으로 나눈다. 그 중 9개를 학습용 자료, 나머지 하나는 검증용 자료로 설정한다.

Step 2. 선택된 SNP 조합에서 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 실험군-대조군의 도수를 기술한다.

Step 3. 실험군-대조군의 비를 구하여 1보다 크거나 같으면 high-risk, 1보다 작으면 low-risk로 정한다. 예를 들어 실험군 = 3, 대조군 = 5인 경우 그 비는 0.6이므로 그 셀은 low-risk로 정한다.

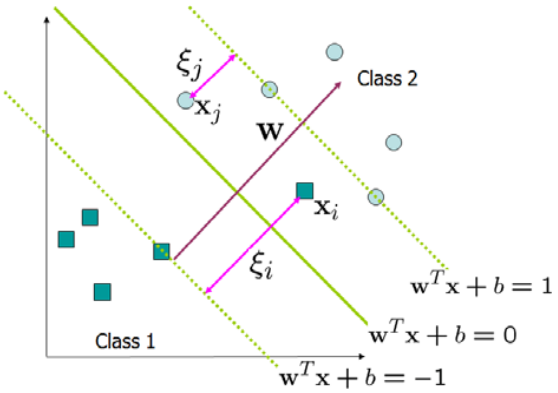

그림 2: *슬랙변수를 도입한 소프트마진기법*

Step 4. 학습용 자료의 모든 셀에서 잘못 분류된 비율인 Misclassification Error(ME)를 구하고, 검증용 자료를 이용하여 잘못 예측된 분류 비율인 Prediction Error(PE)를 구한다.

Step 5. Step 3에서 정의한 10쌍의 모든 데이터셋에 대해서 위의 과정을 반복하여 오분류 오류(ME)의 평균(Avg.ME)과 예측오류(PE)의 평균(Avg.PE), CVC(Cross Validation Consistency) 값을 구한다. CVC 값은 10번의 cross-validation을 시행할 때 각 시행에서 선택된 best model을 카운트하는 것이다 (Chung 등, 2005).

Step 6. 구해진 Avg.ME, Avg.PE, CVC 값을 비교해서 Avg.ME, Avg.PE 값이 낮으며, CVC 값이 큰 조합을 우수 SNP 조합으로 선정한다.

2.2. SVM MDR 방법

SVM MDR 방법은 case-control의 이분형 데이터에만 적용 가능한 MDR 방법의 문제점을 보완하기 위해서 제안된 방법이다 (Lee와 Lee, 2010). 이분형 분류기술인 SVM을 사용하여 연속형 자료를 이분화 함으로써 연속형 자료에서도 MDR 방법을 적용하도록 하는 방법이다. MDR 과정에 선행되어 연속형 목표변수를 입력변수로 하여 SVM 알고리즘을 통해서 Y값이 1인 경우 case, -1인 경우 control로 분류하게 된다. SVM은 Vapnik (1998)에 의해 처음 개발된 통계적 알고리즘으로써 고차원 공간의 데이터를 선형회귀 함수로 재구성함으로써 비선형회귀 문제까지도 해결이 가능한 장점을 가지고 있다. 이러한 장점을 바탕으로 텍스트 분류나, 사진, 영상인식은 물론 생물정보학(Bio information) 분야의 단백질(Protein) 분류, 암(Cancer) 분류에서도 사용되는 효과적인 2 클래스 분류기법이다. SVM의 기본적인 아이디어는 최대 마진 초평면(maximal margin hyperplane)이다. 하지만, 실제의 데이터에서는 이론적인 선형 SVM처럼 완벽한 분류가 이루어지는 경우는 거의 불가능하다. 따라서 최소의 오분류가 존재하게 된다. 이러한 오분류를 최소화시키면서 최대마진(soft margin)을 가지는 초평면을 가지도록 허용하는 슬랙변수(slack variable)를 도입하는 소프트마진기법이 사용된다 (Tan 등, 2006).

이러한 소프트마진기법은 선형으로 완벽히 분리되지 않는 경우에서도 선형의 의사결정 경계를 생성하게 되고 그 목표함수는 다음과 같다.

$$L_p = \frac{1}{2}\|w\|^2 - C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{N}\lambda_i\left(y_i(w_i\cdot x_i+b)-1+\xi_i\right) - \sum_{i=1}^{n}\mu_i\xi_i, \tag{2.1}$$

여기서 C는 규정화 모수(regularization parameter)로 모델 선택 시 고려된다. 위의 목표함수의 KKT조건(Karush-Kuhn-Tucker conditions)은 다음과 같다 (Scholkopf와 Smola, 2001).

$$\epsilon_i > 0, \quad \lambda_i > 0, \quad \mu_i > 0, \tag{2.2}$$

$$\lambda_i[y_i(w \cdot x_i + b) - 1 + \xi_i] = 0, \tag{2.3}$$

$$\mu_i \xi_i = 0. \tag{2.4}$$

이 조건에 의한 이중 라그랑지안은 오분류가 있는 경우와 동일하다. 단, 라그랑지안 승수에 대한 제한 조건들에 차이가 있다. 오분류가 없는 경우는 라그랑지안 승수가 이지만 오분류가 존재하는 경우는 이다. 위의 이중 문제도 제곱프로그래밍 기술을 사용하여 수치적으로 해결 가능하다. 이러한 슬랙변수를 추가하더라도 우리 주위의 비선형분류의 현실문제에서는 효과적이지 못하게 된다. 이러한 비선형의 문제는 커널 함수(Kernel function)를 이용함으로써 해결이 가능하다 (Shim 등, 2009; Cho, 2010). 본래의 좌표 공간에 있는 데이터를 변환할 수 있는 함수가 존재한다고 하자. 그럼 비선형 SVM은 변환된 공간 내에서 다음 형태의 의사결정 경계를 가진다.

$$w \cdot \Phi(x) + b = 0. \tag{2.5}$$

사례 z는 다음의 식을 사용하여 분류될 수 있다.

$$f(z) = \text{sign}\,(w \cdot \Phi(z) + b) = \text{sign}\left(\sum_{i=1}^{n} \lambda_i y_i \Phi(x_i) \cdot \Phi(z) + b\right), \tag{2.6}$$

여기서는 매우 복잡하고 까다로운 벡터 쌍 사이의 내적 계산을 포함한다. 이는 커널 트릭(kernel trick)이란 방법을 통해서 해결할 수 있다. 변환된 공간에서의 두 벡터 사이의 내적은 다음과 같다.

$$\begin{aligned}\Phi(u) \cdot \Phi(v) &= \left(u_1^2, u_2^2, \sqrt{2}u_1, \sqrt{2}u_2, 1\right) \cdot \left(v_1^2, v_2^2, \sqrt{2}v_1, \sqrt{2}v_2, 1\right) \\ &= u_1^2 u_2^2 + v_1^2 v_2^2 + 2u_1 v_1 + 2u_2 v_2 + 1 \\ &= (u \cdot v + 1)^2 .\end{aligned}$$

위의 식은 변환공간의 내적이 원 공간의 유사성함수로 표현됨을 보여준다.

$$K(u, v) = \Phi(u) \cdot \Phi(v) = (u \cdot v + 1)^2 \tag{2.7}$$

함수 K는 원 속성 공간에서의 유사성 함수로써 커널함수로 불린다. 본 논문에서는 사용된 SVM 모형은 RBF(Radial Basis Function)의 커널 함수를 사용하였으며 SPSS사의 Moeler13(구. 클레멘타인)에서 제공되는 디폴트 모수로 RBF 감마 0.1, 규정화 모수는 10으로 설정하였다. 이제 SVM을 통해 연속형 자료가 이분형 형태로 변환되어 MDR 방법의 적용이 가능해지는 것이다. SVM 알고리즘을 통해서 새롭게 생성된 이분형의 목표변수는 하나의 특성뿐 아니라 2개 이상의 특성을 동시에 고려할 수 있게 된다. 즉, 한우의 경제형질을 동시에 고려한 목표변수도 생성이 가능해 지는 것이다.

2.3. D-MDR 방법

목표변수가 연속형일 경우에 MDR 방법의 적용을 위해 제시된 다른 방법으로 더미변수를 활용한 MDR 방법인 D-MDR 방법이 있다. D-MDR 방법은 범주형 독립변수의 경우에 활용하는 더미변수를 활용한 회귀분석을 사용한다. 따라서, 다중의 범주를 더미변수로 변환하여 회귀분석을 진행하고 회

귀분석을 결과로 제시된 회귀계수들의 상호관계를 통해서 양과 음의 상관관계에 따라 high-risk, low-risk로 이분화한다. 2.1절의 Step 3과 4에서 이러한 알고리즘을 적용하게 된다. D-MDR 방법은 오분류율에 대한 평가방법으로 2.1절에서 제시한 오분류율(PE)와 예측오분류율(PE)를 대신해, 평균제곱오차(Average Squared Error; ASE)와 예측평균제곱오차(Prediction Average Squared Error; P-ASE)를 사용한다 (Lee 등, 2009a).

2.4. 순열검정

SVM MDR과 D-MDR 두 방법의 결과 선택된 최적 유전자 조합의 통계적 유의성 검정을 위해서 본 논문에서는 순열검정 (Good, 2000)을 통한 p-value를 계산한다. 순열검정의 절차는 다음과 같다.

Step 1. 가설 설정

- 각 방법을 통해 선별된 우수 유전자형 조합이 특성치에 영향력이 있다.

Step 2. 통계량과 기각역 설정

- 분석에 사용할 통계량으로 F-측정치를 선택한다.
- 특정 유전자형 조합을 가진 그룹과 그 외의 그룹으로 나누어 그룹 간 데이터를 서로 바꾸었을 때 특정 유전자 조합을 가진 그룹의 통계량이 높다면 특성치에 영향력이 있다고 판단한다.

Step 3. 선별된 유전자형 조합에 대한 통계량 계산

- 각 방법을 통해 선별된 우수 유전자형 조합의 F-측정치를 계산한다.

Step 4. 관측치의 재배열과 재배열 후의 통계량 계산

- 두 그룹의 데이터를 n개만큼 랜덤 추출하여 서로 교환한 후 그룹의 F-측정치를 구한다. 이 과정을 10,000번 반복한다.

Step 5. 결론 (p-value 계산)

- 각 F-측정치를 내림차순으로 정렬한 후 기존의 F-측정치와 비교하여 p-value를 구한다.

위의 순열검정을 통해서 선정된 우수 유전자 조합에 대해 p-value를 비교하여 구한다. 3장에서는 SVM MDR과 D-MDR을 실제 한우 데이터에 적용하여 각 방법의 결과를 비교하고, 각 방법의 결과를 통해 한우의 경제형질에 영향을 미치는 우수 유전자 조합을 규명한다.

3. 한우 데이터의 적용결과 비교

3.1. 실험자료

본 연구에서의 데이터는 농협중앙회 한우개량사업소의 후대검정집단인 30차에서 35차 국가 후대검정우 집단 476두로 구성된다. 한우의 여러 경제형질인 일당증체량(ADG; average daily gain)과 도체중(CWT; carcass cold weight), 등심단면적(LMA; longissimus muscle dorsi area), 근내지방도(MS; marbling score)는 모든 F1 자손으로부터 수집되었고, 한국 축산물등급판정소의 규격에 따라 측정되었다. EST based SNP 연관지도 (Snelling 등, 2005)에서 후보 유전자로 판단되는 LOC534614 유전자내 17개의 SNPs 중 6개의 htSNP(haplotype-tagging SNP)를 한우의 경제형질에 영향을 주는 후보 SNPs로 판단한다 (Lee, 2009b). 여기서 한우의 경제형질은 등급을 결정하는 요인으로 일당증체량, 도체중량, 등심단면적, 근내 지방도이다.

표 1: 종합 경제형질에 대한 SVM MDR 적용결과

요인의 수	최적 유전자 조합	Avg.ME	Avg.PE	CVC	순열검정(p-value)	
1	g.11500-117C>G	0.4398	0.4398	10	ADG	0.0186
					MS	0.0008
					CWT	0.5081
					LMA	0.5178
2	g.8778G>A g.11500-117C>G	0.4018	0.4018	10	ADG	0.0153
					MS	0.0008
					CWT	0.0127
					LMA	0.1669
3	g.8778G>A g.11500-117C>G g.66995-169insdelC	0.3627	0.3666	10	ADG	<0.001
					MS	<0.001
					CWT	0.0012
					LMA	0.3432

표 2: 일당증체량(ADG)에 대한 D-MDR 적용결과

요인의 수	최적 유전자 조합	일당증체량			
		ASE	P-ASE	CVC	순열검정(p-value)
1	g.34425+102A>T	0.007778	0.007740	10	0.1476
2	g.8778G>A g.11500-117C>G	0.007821	0.007775	10	0.0108
3	g.8778G>A g.11500-117C>G g.66995-169insdelC	0.007524	0.007479	10	0.0206

3.2. SVM MDR 방법의 적용 결과

다음의 표 1은 SVM MDR을 한우 데이터에 적용시킨 결과이다. class 변수로는 한우의 종합 경제형질인 일당증체량(ADG), 도체중량(CWT), 등심단면적(LMA), 근내지방도(MS)를 동시에 고려한 값으로써 4개의 연속형 변수인 경제형질을 입력변수로 받아 SVM 알고리즘을 통해서 이분화한 값을 사용하였다.

10번의 cross-validation 결과 나온 평균오분류오차(Avg.ME)와 평균예측오분류오차(Avg.PE)을 구하였다. 그리고 각 시행에서 선택된 우수 유전자조합의 횟수인 CVC 값을 기준으로 최종적인 우수 유전자 조합을 선별하였다. 2개의 조합에서는 g.8778G>A, g.11500-117C>G의 조합이 Avg.ME와 Avg.PE가 0.4018, CVC 값이 10으로 우수 유전자 조합으로 선택되었으며, 3개의 조합에서는 g.8778G>A, g.11500-117C>G, g.66995-169insdelC의 조합이 Avg.ME가 0.3627, Avg.PE가 0.3666, CVC 값이 10으로써 최종 우수 유전자조합으로 선택되었다. 선택된 최적 유전자 조합의 통계적 유의성을 확인하기 위한 순열검정에서 등심단면적(LMA)를 제외한 나머지 3가지 경제형질에서는 모두 표 1과 같이 유의성을 보였다. 표 1에서 등심단면적의 경우는 순열검정의 결과가 모두 유의하지는 않았지만 단일 SNP의 경우보다 상호작용에 의한 경우 그 값이 많이 낮아지는 것을 확인할 수 있었다. 이는 특성치에 대한 영향이 단일 효과보다는 상호작용의 효과에 더 큰 영향을 받는 것을 의미한다.

3.3. D-MDR 방법의 적용 결과

위의 표 2와 표 3은 한우 데이터에 D-MDR 방법을 적용시킨 결과이다. D-MDR 방법에서는 종합 경제형질을 한 번에 적용할 수 없다. 따라서 각각의 경제형질에 대해서 D-MDR 결과를 적용한 다음 공통적인 요인을 선택하여 그 조합을 우수 유전자조합으로 선택하게 된다. 본 연구에서는 공통적인

표 3: 근내지방도(MS)에 대한 D-MDR 적용결과

요인의 수	최적 유전자 조합	일당증체량			
		ASE	P-ASE	CVC	순열검정(p-value)
1	g.34425+102A>T	15.0710	15.0247	10	0.0097
2	g.8778G>A g.11500-117C>G	15.0184	14.9371	10	0.0150
3	g.8778G>A g.11500-117C>G g.34425+102A>T	14.8893	14.8357	10	0.0153

표 4: 한우 종합경제형질에 영향을 주는 최적 유전자 조합

SVM MDR			D-MDR		
경제형질		최적 유전자 조합	경제형질		최적 유전자 조합
ADG MS CWT LMA	2개 조합	g.8778G>A g.11500-117C>G	ADG	2개 조합	g.8778G>A g.11500-117C>G
				3개 조합	g.8778G>A g.11500-117C>G g.66995-169insdelC
	3개 조합	g.8778G>A g.11500-117C>G g.66995-169insdelC	MS	2개 조합	g.8778G>A g.11500-117C>G
				3개 조합	g.8778G>A g.11500-117C> g.34425+102A>T

요인을 찾기 위해서 한우 경제형질 4가지를 대상으로 D-MDR 방법을 적용하였다. 하지만 4개의 경제형질에 대한 D-MDR 적용결과에서 공통의 유전자를 찾는 것은 쉽지 않다. D-MDR의 결과에서도 일당증체량과 근내지방도에서는 공통의 유전자 조합을 확인할 수 있었지만 도체중량과 등심단면적의 결과에서는 공통의 결과를 찾을 수 없었다. 따라서 우리는 SVM MDR과의 비교를 위해서 D-MDR의 결과 중 일당증체량과 근내지방도의 결과를 바탕으로 한우의 육질과 양의 등급을 높일 수 있는 최적의 유전자 조합을 규명하였다. 먼저 표 2는 한우의 경제형질 중 일당증체량에 대해서 D-MDR 방법을 적용한 결과이다. 일당증체량에 대한 결과는 SVM을 활용한 MDR과 동일한 우수 유전자 조합이 선택되었다. 먼저 2개의 조합에서 g.8778G>A, g.11500-117C>G의 조합이 ASE가 0.007821, P-ASE가 0.007775, CVC 값이 10으로 우수 유전자 조합으로 선택되었으며, 이 조합에 대한 순열검정의 결과에서도 p값이 0.0108로써 이들 조합이 일당증체량에 영향을 준다는 것을 입증하고 있다. 3개의 조합에서 g.8778G>A, g.11500-117C>G, g.66995-169insdelC의 조합이 ASE값이 0.007524, P-ASE 값이 0.007479, CVC 값은 10으로 우수 유전자 조합으로 선택되었고, 이에 대한 순열검정의 p값도 0.0206으로 일당증체량에 유의한 영향을 주는 것을 알 수 있다. 근내지방도에 대한 D-MDR 결과를 살펴보면 2개의 조합에서는 SVM을 활용한 MDR과 D-MDR과 동일하게 g.8778G>A, g.11500-117C>G의 조합이 ASE가 15.0184, P-ASE가 14.9371, CVC 값이 10으로 우수 유전자 조합으로 선택되었으며, 순열검정에 대한 p값은 0.0150으로 나타났다. 그리고 3개의 조합에서는 g.8778G>A, g.11500-117C>G, g.34425+102A>T의 조합이 ASE는 14.8893, P-ASE는 14.8357, CVC 값이 10으로 우수 유전자 조합으로 선택되었다. 이에 대한 순열검정의 p값은 0.0153으로써 근내지방도에 대한 D-MDR 방법의 결과 선택된 우수 유전자 조합이 모두 근내지방도에 유의한 영향을 미치는 것으로 나타났다. SVM을 활용한 MDR 방법과 D-MDR 방법에 의한 결과를 정리하면 표 4와 같이 나타난다.

각 조합에서 g.8778G>A, g.11500-117C>G 조합이 모두 공통적으로 포함된 것을 알 수 있다. 따

라서 우리는 SVM을 활용한 MDR 방법과 D-MDR 방법의 결과를 종합해본결과 g.8778G>A, g.11500-117C>G의 조합이 한우의 일당증체량과 근내지방도를 포함한 종합경제형질에 영향을 주는 우수 유전자 조합으로 최종 선택하였다.

4. 결론 및 토의

복잡한 유전학에서 인간의 질병과 가축의 특성에 영향을 주는 요인을 찾는 연구가 가운데 그것들의 상호작용에 대한 연구에 더 큰 비중을 두고 있다. 단일 유전자의 효과보다는 유전자와 유전자간의 상호작용으로 인한 효과가 더욱 크게 작용하고 있다는 말이다. 그 대표적인 예로써 MDR 방법이 제시되었다. 하지만 MDR 방법은 2클래스의 이분형 자료에만 적용이 가능하다는 한계가 있다. 그러한 한계를 극복하기 위해서 E-MDR, D-MDR과 같은 방법들이 새롭게 제시되었고 최근 연구에서는 2클래스 분류기술인 SVM 알고리즘을 활용하여 연속형 자료를 이분화시킴으로써 MDR 방법이 적용가능하도록한 SVM을 활용한 MDR 방법도 제시되었다. D-MDR 방법은 기존의 MDR 방법내에서 High-risk 그룹과 Low-risk 그룹을 분류하는 과정에서 더미변수를 활용하기 때문에 원래 데이터를 그대로 사용해도 된다는 장점이 있다. 그러나 한우의 경제형질과 같은 자료와 같이 목표변수를 동시에 고려해야할 경우는 목표변수의 수만큼의 절차를 반복해서 공통적인 유전자 조합을 최종의 유전자 조합으로 선택하여야 한다. 또한 공통의 우수 유전자 조합을 찾지 못하는 경우도 발생한다. 하지만 SVM을 이용한 MDR 방법에서는 목표변수를 입력변수로 하는 SVM 알고리즘을 통해서 하나의 새로운 목표변수를 생성하게 된다. 이러한 방법은 기존 방식의 절차를 한번의 절차로 줄일 수 있다. 새롭게 생성된 변수를 바탕으로 최적의 유전자 조합을 찾기 때문에 한 번에 최적의 유전자 조합을 선별할 수 있다는 장점이 있다. 우리는 본 연구에서 각 SVM을 활용한 MDR 방법과 D-MDR 방법을 통해서 한우의 종합 경제 형질에 영향을 주는 우수 유전자 조합을 찾아내고, 두 방법의 결과를 가지고 최종적으로 한우의 종합 경제 형질에 영향을 주는 우수 유전자 조합을 규명하였다. 그 결과 SVM을 활용한 MDR 방법과 D-MDR 방법에서 모두 공통적으로 선택된 조합인 g.8778G>A, g.11500-117C>G의 조합을 한우의 종합 경제형질에 영향을 주는 우수 유전자 조합으로 선택하였다. 그리고 선택된 위 조합의 통계적 유의성을 입증하기 위한 순열검정 결과에서도 모두 유의한 것으로 나타났다. 이러한 연구 결과를 바탕으로 위의 유전자형을 개량하게 된다면 일당증체량이나 근내지방도를 포함한 한우의 종합경제형질에서 보다 더 가치있는 한우를 개발할 수 있을 것이다. 아울러 제시한 SVM을 활용한 MDR 방법이나 D-MDR 방법의 각 장단점을 잘 고려해서 기존의 데이터를 살리면서 SVM 알고리즘을 적용할 수 있는 S-MDR(가명)과 같은 방법도 연구될 수 있을 것으로 생각된다.

참고 문헌

Cho, D. (2010). Mixed-effects LS-SVM for longitudinal data, *Journal of Korean Data & Information Science Society,* **21**, 363–369.

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

Good, P. (2000). Permutation Test: A practical guide to resampling method for testing hypotheses, *Springer-Verlag Berlin and Heidelberg GmbH & Co.*, New York.

Lee, H. G. (2009a). Power of multifactor dimensionality reduction with dummy variable and detecting best gene interaction, *M.S. Thesis*, 1–53.

Lee, J. Y. and Lee, J. H. (2010). Support vector machine and multifactor dimensionality reduction for detecting major gene interactions of continuous data, *Journal of Korean Data & Information Science Society*, **21**, 1271–1280.

Lee, Y. S. (2009b). *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, 1–94.

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.

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press.

Shim, J., Park, H. and Seok, K. H. (2009). Variance function estimation with LS-SVM for replicated data, *Journal of Korean Data & Information Science Society,* **20**, 925–931.

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, *BioMed Central Genomics,* **6**, 74–84.

Vapnik, V. (1998). *Statistical Learning Theory*, John Wiley & Sons, New York.

Tan, P., Steinbach, M. and Kumar, V. (2006). *Introduction to Data Mining*, Addison-Wesley, New-York.

2011년 1월 접수; 2011년 5월 채택

A Comparison Study on SVM MDR and D-MDR for Detecting Gene-Gene Interaction in Continuous Data

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

[a]Department of Statistics, Yeungnam University

Abstract

We have used a multifactor dimensionality reduction(MDR) method to study the major gene interaction effect in general; however, without application of the MDR method in continuous data. In light of this, many methods have been suggested such as Expanded MDR, Dummy MDR and SVM MDR. In this paper, we compare the two methods of SVM MDR and D-MDR. In addition, we identify the gene-gene interaction effect of single nucleotide polymorphisms(SNPs) associated with economic traits in Hanwoo(Korean cattle). Lastly, we discuss a new method in consideration of the advantages that the other methods present.

Keywords: Gene-gene interaction, MDR method, SNP, SVM algorithm.

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