Journal of the Korean Data & Information Science Society
2010, **21**(6), 1271–1280

한국데이터정보과학회지

서포트 벡터 머신 알고리즘을 활용한 연속형 데이터의 다중인자 차원축소방법 적용

이제영[1] · 이종형[2]

[12]영남대학교 통계학과

접수 2010년 10월 8일, 수정 2010년 11월 18일, 게재확정 2010년 11월 23일

요 약

인간의 질병과 가축의 특성에 영향을 주는 유전자들의 상호작용을 규명하는 방법으로 전통적인 통계방법들이 사용되었지만, 유전자와 같은 고차원의 데이터에는 적합하지 않았다. 따라서 다중인자 차원축소방법이 제안되었다. 다중인자 차원축소방법은 모형에 대한 가정이 필요하지 않는 비모수적 방법으로 이분형 자료에 적용 가능 하지만, 연속형 데이터에는 적용할 수 없는 단점이 있다. 따라서 본 연구에서는 일반화 분류 성능이 뛰어난 서포트 벡터 머신 알고리즘을 통해 연속형 자료를 가공하여 다중인자 차원축소방법에 적용하였다. 아울러 한우의 6번 염색체내 6개의 후보 단일염기다형성을 대상으로 연속형 자료인 실제 한우의 경제형질에 서포트 벡터 머신을 이용한 다중인자 차원축소방법을 적용함으로써 한우의 경제형질에 연관된 우수 유전자 상호작용의 조합을 규명하였다.

주요용어: 다중인자 차원축소방법, 단일염기다형성, 서포트 벡터 머신, 한우 경제형질.

1. 서론

현대 유전학에서의 관심사 중 하나는 바로 인간의 질병 및 가축의 특성과 연관된 유전자를 규명하는 것이다. 일반적으로 인간의 질병과 가축의 특성은 단일 유전자의 영향보다는 유전자-유전자 간의 상호작용으로 일어난다고 믿고 있으며, 그러한 이유에서 많은 연구들이 유전자간의 상호작용 규명을 위한 모형으로 표준 통계적 모형을 사용해왔다. 이러한 모형은 변수가 늘어남에 따라 모형의 복잡도가 증가되거나, 해석이 어려워져 실제로 해석 불가능한 경우도 종종 발생을 한다. 또한, 모형화 된 경우라도 가능한 테이블 셀에 관측값이 없는 상황이 발생하게 된다. 이처럼 많은 수의 인간의 유전자를 고려했을 경우 이제까지 사용된 표준 통계적 모형은 한계점에 부딪히게 된다. 그래서 제안된 방법이 다중인자 차원축소 (Multifactor dimensionality reduction, MDR; Richie 등, 2001)이다. MDR 방법은 상호작용을 규명하기 위한 비모수적 방법으로써, 모형에 대한 어떠한 가정도 필요하지 않다. 적당한 차수의 데이터로 축소함으로써 변수들 사이의 복잡한 관계까지 밝힐 수 있다. 하지만 실험-대조의 이분형 자료에만 적용이 가능할 뿐, 연속형 자료에는 적용할 수 없는 단점이 있다. 본 논문에서 우리는 서포트 벡터 머신 (Support Vector Machine; SVM) 알고리즘을 사용하여 연속형 데이터를 MDR 방법에 적용한다. SVM은 Vapnik (1998)에 의해 처음 개발된 통계적 알고리즘으로써 구조적 위험최소화 (Structural Risk Minimization, SRM)를 통해 오류를 최소화 시키는 방법에 토대를 두고 있다 (Cristianini 등, 2000). 최근 상당한 주목을 받은 분류 기술 중 하나로써, 손으로 쓴 숫자인식에서부터 텍스트, 영상기술

[1] 교신저자: (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 교수. E-mail: jlee@yu.ac.kr
[2] (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 석사과정.

에 이르기까지 많은 실제 응용에서 유망한 결과를 보여주었다 (Lim 등, 2010). 또한 고차원데이터에서도 잘 적합할 수 있어 차원 문제를 피해 갈 수 있다. 이와 같은 SVM의 특성을 고려해 연속형 데이터를 변환함으로써 MDR방법의 적용이 가능해 지는 것이다. 본 논문의 구성은 다음과 같다. 제 2장에서는 본 연구에서 사용된 분류 기술인 SVM 알고리즘에 대해서 소개한다. 아울러 유전자의 상호작용 규명을 위한 방법인 MDR방법에 대해서 간단하게 소개한다. 제 3장에서는 SVM 알고리즘을 통해서 연속형 데이터가 어떻게 MDR에 적용되는가에 대해서 자세히 묘사하고자 한다. 그리고 마지막 제 4장에서는 실제 한우데이터를 위의 프로세스에 적용함으로써 한우의 경제형질에 연관된 우수 유전자 조합을 규명하고, 끝으로 향후 연구에 대해 고찰한다.

2. SVM의 소개와 MDR의 확장

2.1. 기본적인 SVM의 원리

SVM은 고차원 공간의 데이터를 선형회귀 함수로 재구성함으로써 비선형회귀 문제까지도 해결이 가능한 장점을 가지고 있다. 이러한 장점을 바탕으로 텍스트 분류나, 사진, 영상에서의 인식은 물론 Bio informatics 분야의 Protein 분류, Cancer 분류에서도 사용되는 효과적인 2 클래스 분류기법이다. SVM의 기본적인 아이디어는 최대 마진 초평면 (maximal margin hyperplane) 이다.

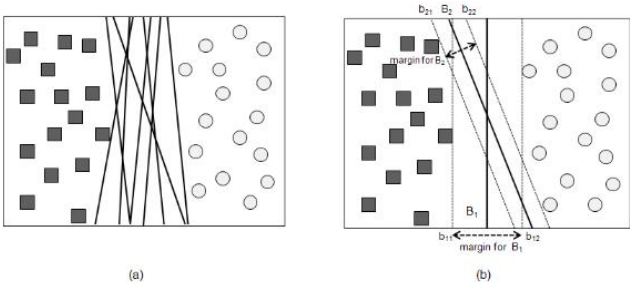

그림 2.1 의사 결정 경계와 마진

그림 2.1의 (a)는 네모와 원으로 표현된 다른 두 집단이 선형으로 분류되어 있다. 즉 네모들은 초평면의 왼쪽, 모든 원들은 초평면의 오른쪽에 위치하고 있는 것이다. 하지만 (a)에서 보듯이 그러한 초평면들은 무수히 많이 존재한다. 초평면이 일반화 오류와 어떤 연관이 있는지는 (b)의 그림을 보면 알 수 있다. 두 개의 의사결정 경계인 초평면 B_1과 B_2가 있다. 여기서 B_{i1}, B_{i2}는 의사결정 초평면을 네모, 원에 닿을 때까지의 평행 이동시켜 놓은 것이다. 바로 B_{i1}, B_{i2}사이의 거리를 분류기의 마진이라 부른다. SVM 알고리즘의 핵심은 이러한 마진을 최대화 시키는 초평면을 찾는 것이다 (Cho, 2010).

의사결정 초평면은 다음과 같은 식으로 표현된다.

$$W \cdot X + b = 0$$

여기서 시험 사례 z에 대한 클래스 레이블 y는 다음과 같이 예측할 수 있다.

$$y = \begin{cases} 1, & w \cdot z + b > 0 \\ -1, & w \cdot z + b < 0 \end{cases}$$

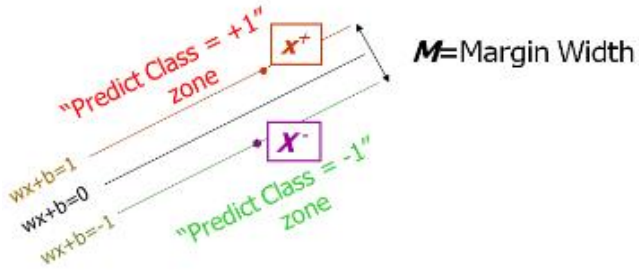

그림 2.2 SVM을 이용한 연속형 데이터의 MDR 적용과정

위의 식에서 보듯이 클래스 $y = 1$의 모든 훈련 사례는 의사결정 경계 위에 위치하고 클래스 $y = -1$의 모든 훈련사례는 의사결정 경계 아래에 위치해야 한다. 여기서 의사결정 경계의 두 인자 w와 b를 조정하면, 그림2.2와 같이 의사결정 경계에 평행한 두 개의 초평면을 얻을 수 있다. 여기서 이 두 초평면사이의 거리가 바로 마진 (M)이 되는 것이다. 마진 (M)은 한 평면에서 다른 평면을 빼면 구할 수 가 있다.

$$w \cdot (x_1 - x_2) = 2||w|| \times M = 2$$

$$\therefore M = \frac{2}{||w||}$$

여기서 마진의 최대화는 아래 목표 함수의 최소화와 동등하다.

$$f(w) = \frac{||w||^2}{2}$$

목표 함수가 제곱에 비례하고, 조건들은 매개변수 w와 b의 선형으로 비례하기 때문에 이른 컨벡스 (convex) 최적화 문제라 부르며 표준 라그랑즈 승수 (Lagrange multiplier) 기법을 사용하여 해결 할 수 있다.

$$L_p = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} \lambda_i (w_i \cdot x_i + b) - 1)$$

라그랑지안을 최소화하기위해서 L_p를 w와 b에 대해 편미분해서 풀어야 한다.

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{N} \lambda_i y_i x_i$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \lambda_i y_i = 0$$

하지만, 라그랑즈 승수의 값을 모르기 때문에 w, b에 대해서 해를 구할 수가 없다. 따라서 이들을 등식 조건들의 집합으로 변환시키게 되는데 이것이 KKT (Karush-Kuhn-Tucker) 조건이다 (Schölkopf 등, 2001).

$$\lambda_i \geq 0 \lambda_i [y_i (w \cdot x_i + b) - 1] = 0$$

이제 위의 최적화 문제는 라그랑지안을 라그랑즈 승수들만의 함수로 변환하여 간소화 시키는 dual problem이 되고 다음과 같은 이중형태로 표현된다.

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

이러한 이중 최적화 문제는 제곱프로그래밍과 같은 기술에 의해서 해결 가능하다. 따라서 최종적인 결정 경계는 다음과 같이 표현된다.

$$(\sum_{i=1}^{N} \lambda_i y_i x_i \cdot x) + b = 0$$

하지만, 실제로 앞에서 설명한 SVM처럼 완벽한 분류가 이루어지는 경우는 거의 불가능하다. 따라서 어느 정도의 오분류가 존재하더라도, 최대 마진을 가지는 초평면을 가지도록 허용하는 슬랙변수 (slack variable)를 추가하는 소프트 마진 (soft margin)기법이 필요하다 (Tan 등, 2006) 이 방법은 클래스가 선형으로 분리되지 않는 상황에서도 선형의 의사결정 경계를 생성할 수 있다. 변경된 목표 함수는 다음과 같다.

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

C는 규정화 모수 (regularization parameter)로써 유효집합에서의 모델 성능에 따라 선택된다. 여기서의 KKT조건은 다음과 같다.

$$\xi_i > 0, \lambda_i > 0, \ \mu_i > 0$$
$$\lambda_i [y_i (w \cdot x_i + b) - 1 + \xi_i] = 0$$
$$\mu_i \xi_i = 0$$

이 조건에 의한 이중 라그랑지안은 오분류가 있는 경우와 동일하다. 단, 라그랑지안 승수에 대한 제한조건들에 차이가 있다. 오분류가 없는 경우는 라그랑지안 승수가 $\lambda_i \geq 0$이지만 오분류가 존재하는 경우는 $0 \leq \lambda_i \leq C$이다. 위의 이중 문제도 제곱프로그래밍 기술을 사용하여 수치적으로 해결 가능하다.

2.2. 비선형 SVM

우리는 2.1절에서 각 클래스의 분류방법으로 선형 SVM의 기본원리를 소개하고 의사결정경계를 구축하였다. 하지만 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). 이 절에서는 위와 같은 문제 해결을 위한 비선형의 의사결정 경계를 갖는 자료에 SVM을 적용하는 방법을 기술한다. 본래의 좌표 공간 x에 있는 데이터를 선형 의사결정 경계를 사용할 수 있는 공간으로 변화시키는 것이다. 변환 후의 과정은 앞 절의 방법을 적용 할 수 있다. 여기서 주어진 데이터를 적절하게 변환할 수 있는 함수 $\Phi(x)$가 존재한다고 하자. 그럼 비선형 SVM은 변환된 공간 내에서 다음 형태의 의사결정 경계를 가진다.

$$w \cdot \Phi(x) + b = 0$$

2.1절에서의 선형 SVM의 방법을 적용하여 제한된 최적화 문제에 대한 다음의 이중 라그랑지안을 유도할 수 있다.

$$L_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \Phi(x_i) \cdot \Phi(x_j)$$

마지막으로 시험 사례 z는 다음의 식을 사용하여 분류될 수 있다.

$$f(z) = sign(w \cdot \Phi(z) + b) = sign(\sum_{i=1}^{n} \lambda_i y_i \Phi(x_i) \cdot \Phi(z) + b)$$

여기서는 벡터 쌍 사이의 내적 계산을 포함한다. 이러한 계산은 매우 까다로우며, 고차원 문제의 저주에 처할 수 있는데, 이는 커널 트릭 (kernel trick)이란 방법을 통해서 해결할 수 있다. 커널 트릭은 원래의 속성 집합을 사용하여 변환된 공간에서 유사성을 계산하는 방법이다. 변환된 공간에서의 두 벡터 u와 v사이의 내적은 다음과 같다.

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

위의 식은 변환공간의 내적이 원 공간의 유사성함수로 표현됨을 보여준다.

$$K(u, v) = \Phi(u) \cdot \Phi(v) = (u \cdot v + 1)^2$$

함수 K는 원 속성 공간에서의 유사성 함수로써 커널함수로 불린다. 다음의 식은 커널 함수를 사용하여 SVM에 의해 구해진 비선형 의사결정 경계를 보여준다.

$$\begin{aligned} f(z) &= sign(\sum_{i=1}^{n} \lambda_i y_i \Phi(x_i) \cdot \Phi(z) + b) \\ &= sign(\sum_{i=1}^{n} \lambda_i y_i K(x_i, z) + b) \\ &= sign(\sum_{i=1}^{n} \lambda_i y_i (x_i \cdot z + 1)^2 + b \end{aligned}$$

이러한 커널 트릭은 비선형 SVM의 문제점들을 해결 해 준다. 먼저 커널함수가 머서의 정리 (Mercer's therem)란 수학 원리를 만족하기 때문에 변환 함수의 정확한 형태를 알 필요가 없을뿐더러 내적의 계산이 변환 함수 $\Phi(x)$를 사용하는 것보다 훨씬 빠르다. 그리고 계산 자체가 원래의 공간에서 이루어지므로 고차원 문제를 피해 갈 수 있는 것이다. 일반적으로 비선형 SVM의 커널 함수를 이용한 변환이 많이 이용된다. 본 논문에서는 사용된 모델은 오분류가 있는 비선형 SVM으로 SPSS사 Modeler13 프로그램에서 RBF (Radial Basis Function)의 커널 함수를 사용하였으며 제공되는 디폴트 모형으로 RBF 감마 0.1 규정화 모수는 10으로 설정하였다.

2.3. MDR 방법

Multifacrot Dimension Reduction (MDR) 방법은 일반화된 선형 모형의 전통적 통계기법과는 달리 모수에 대한 추정과 모형에 대한 가정이 필요하지 않은 비모수적 방법이다. Richie 등 (2001)이 제안한 방법으로 실험-대조의 이분형 자료에 적합가능하다. 또한 MDR방법은 검정력 평가를 통해서 높은 검정력을 보인 우수한 유전자 상호작용 효과 규명을 위한 방법이다. 하지만, 이분형 자료에만 적용이 가능할 뿐 연속형 자료에는 적용을 할 수 없다는 단점이 있다. 이러한 MDR방법의 단점을 보완하기 위해서 많은 연구들이 진행되었다. CART알고리즘을 활용한 다중인자 차원축소방법 (Lee 등, 2008)은 데이터마이닝 방법의 CART알고리즘을 활용하여 연속형 자료를 이분화하는 방법으로 Lee 등에 의해 검정력이 확인되었다 (Lee 등, 2010). 또한 더미변수를 활용한 다중인자 축소방법 (D-MDR; Lee 등, 2009)방법 역시 연속형 자료를 MDR방법의 단점을 보완한 방법이다. 이처럼 MDR과 관련된 많은 연구들이 현재도 진행되고 있다. 다음장에서는 앞에서 소개한 SVM 알고리즘을 통해 연속형 자료의 MDR방법 적용 과정을 설명한다.

3. 서포트 벡터 머신의 적용

3.1. SVM과 MDR의 적용방법

앞서 우리는 높은 분류성능을 가진 SVM알고리즘을 소개하였다. 특히 2.1절과 2.2절에서 소개한 SVM은 최근 많은 분야에서 이용될 만큼 이분형 분류 성능이 뛰어나다. 또한 SVM 알고리즘은 Modeler13의 응용프로그램을 통해 쉽게 적용가능하다. 2.3절에서 소개한 MDR방법은 유전자 상호작용 효과를 규명하는 방법이지만 이분형 자료에만 적용 가능하다는 단점이 있다. 이 장에서는 연속형 데이터를 MDR에 적용시키기 위해서 SVM알고리즘을 이용하고자 한다. 높은 이분형 분류 성능을 가진 SVM의 특징을 이용해 연속형 데이터를 변환하여, MDR에 적용하고자 하는 것이다. 간단한 절차는 다음과 같다.

Step 1. 관심의 대상이 되는 연속형 변수들을 입력변수로 설정한다.

Step 2. SVM 알고리즘을 이용해 step 1에서의 연속형 변수를 입력받아 새로운 이분형 변수 $Y = \begin{cases} 1, & high-group \\ -1, & low-group \end{cases}$ 를 생성한다. 여기서 Y값이 1인 경우 실험군, -1인 경우 대조군으로 한다.

Step 3. 데이터를 크기가 동일한 10개의 셋으로 나눈다. 그중 9개를 학습용 자료, 나머지 하나는 검증용 자료로 설정한다.

Step 4. 선택된 SNP조합에서 SNP의 각 수준을 기초로 한 개체들을 multifactor classes 또는 cells에 기술한다. 예를 들어서 k=2일 경우, SNP는 3개의 수준으로 되어있으므로 $3^2 = 9$개의 셀을 가진다. 각각 9개의 셀에 실험군-대조군의 도수를 기술한다.

Step 5. 실험군-대조군의 비를 구하여 1보다 크거나 같으면 high-risk, 1보다 작으면 low-risk로 정한다. 즉, 실험군=3, 대조군=5 인 경우 0.6으로 low-risk로 정한다.

Step 6. 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 7. 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 카운

트 하는 것이다 (Chung 등, 2005).Step 8. 구해진 Avg.ME, Avg.PE, CVC 값을 비교해서 Avg.ME, Avg.PE 값이 낮으며, CVC 값이 큰 조합을 우수 SNP 조합으로 선정한다.

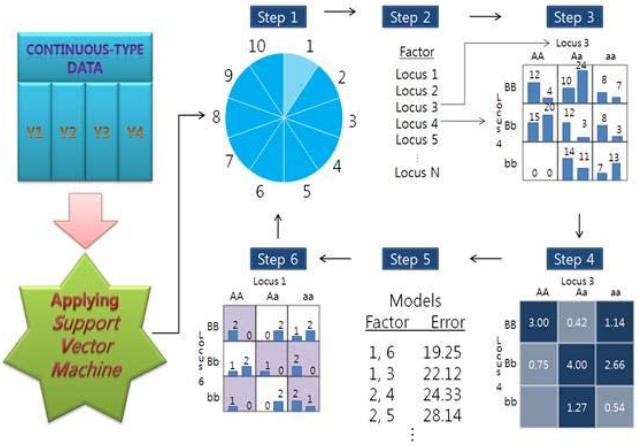

그림 3.1 SVM을 이용한 연속형 데이터의 MDR 적용과정

다음 절에서는 실제 한우데이터 위의 절차를 적용하고자 한다. 개체수의 보완을 위해 Bootstrap (Eforn 등, 1993)방법을 사용하여 개체수의 균형을 맞춰 우수한 상호작용의 효과를 지닌 유전자 조합을 선별한다.

3.2. 한우 데이터의 적용

본 연구의 데이터는 농협중앙회 한우개량사업소의 후대검정집단인 30 35차 국가 후대검정우 집단 476두로 구성된다. 한우의 경제형질인 일당증체량 (ADG)과 도체중량 (CWT), 등심단면적 (LMA), 근내지방도 (MS)의 4개의 연속형 변수에 영향을 주는 우수 유전자 조합을 찾기 위해 6개의 htSNP (haplotype-tagging SNP; Lee, 2009)를 사용하여 3.1절의 절차를 적용하였다. 그 결과 한우의 종합적인 경제형질에 영향을 주는 우수 유전자 조합을 표 3.1과 같이 규명하였다.

표 3.1 SVM변환된 연속형 데이터의 우수 유전자 조합 MDR 적용결과

SNP 조합의 수	우수 유전자 조합 결과			
	최적 조합	Avg.ME	Avg.PE	CVC
1	g.11500-117C>G	0.4398	0.4398	10
2	g.8778G>A g.11500-117C>G	0.4018	0.4018	10
3	g.8778G>A g.11500-117C>G g.66995-169insdelC	0.3627	0.3666	10

위의 표 3.1은 한우의 경제형질에 해당하는 연속형 데이터인 일당증체량 (ADG)과 도체중량 (CWT), 등심단면적 (LMA), 근내지방도 (MS)을 입력변수로 하는 SVM알고리즘을 통해서 생성된 값을 class변수로 MDR방법을 적용한 결과이다. 10번의 cross-validation 결과 나온 오분류 오류 (ME)와 예측오류 (PE)값을 평균하여 Avg.ME와 Avg.PE을 구하였다. 그리고 각 시행에서 선택된 우수 유전자조합

의 횟수인 CVC값을 기준으로 최종적인 우수 유전자 조합을 선별하였다. 일당증체량 (ADG)과 도체중량 (CWT), 등심단면적 (LMA), 근내지방도 (MS) 4가지 경제형질을 동시에 고려한 유전자의 상호작용 중 2개의 조합에서는 g.8778G>A g.11500-117C>G의 조합이 Avg.ME와 Avg.PE가 0.4018로써 최고의 모형으로 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. 단일 효과에서는 g.11500-117C>G가 가장 많은 영향을 주는 유전자로 선별되었다.

4. 결론

인간과 관계된 복합질병 및 가축의 경제특성에 관련된 유전자의 상호작용 효과를 규명하기 위한 방법으로 MDR방법이 제시되었다. 전통적인 방법의 통계적 방법들은 모형에 대한이 필요하지만, MDR방법은 모형에 대한 가정이 필요하지 않은 비모수 방법으로 검증된 높은 검정력을 가지는 우수한 방법이다. 하지만 이분형 자료에만 적용이 가능하다는 단점이 있다. 따라서, 우리는 본 연구에서 연속형 자료를 MDR방법에 적용시키고자 SVM 알고리즘을 소개하였다. SVM알고리즘은 높은 분류성능을 가진 분류알고리즘으로써 특히 이분형 자료의 경우게 일반화의 효율이 높은 것이 특징이다. 이러한 특징을 바탕으로 우리는 한우의 경제형질인 연속형 자료의 일당증체량 (ADG)과 도체중량 (CWT), 등심단면적 (LMA), 근내지방도 (MS)의 연속형 자료를 이분화함으로써 MDR방법에 적용이 가능해진다. SVM 알고리즘을 통해서 변환된 변수를 MDR방법에 적용한 결과 한우의 경제형질에 연관된 유전자 상호작용의 효과를 규명할 수 있었다. 그 결과 2개 요인의 상호작용에서는 g.8778G>A g.11500-117C>G 의 SNP조합이 최고 우수 유전자 조합으로 선별되었고, 3개의 요인에 대한 상호작용의 효과에서는 g.8778G>A g.11500-117C>G g.66995-169insdel의 SNP조합이 가장 우수한 유전자 조합으로 선별되었다. 따라서, 우리는 MDR방법에 적용할 수 없었던 연속형 자료를 SVM 알고리즘을 활용해 MDR방법에 적용함으로써 MDR방법의 한계를 극복할 수 있었다. 아울러 한우의 경제형질과 관련된 우수 유전자 상호작용의 조합을 규명할 수 있었다. 본 연구는 앞서 2장에서 언급했던 E-MDR이나 D-MDR과 같은 방법과 마찬가지로 연속형 자료를 MDR방법에 적용할 수 있을 뿐 아니라, 기존의 방법에서 하나의 경제형질에 대해서 각각의 우수 유전자 조합을 선별하고, 공통적인 조합을 찾는 방법의 번거로움이 SVM 알고리즘을 활용할 경우 모든 경제형질을 한 번에 고려하게 되므로 그만큼 절차가 간단해지고 쉬워졌다. 추후 연구에서는 기존에 제안되었던 E-MDR이나 D-MDR과의 비교를 통해서 각 방법의 장점과 단점을 바탕으로 더 정확한 선별 도구의 제안도 기대된다.

참고문헌

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

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press.

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

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

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

Lee, J. Y., Lee, J. H. and Lee, H. G. (2010). Power of expanded multifactor dimensionality reduction with CART algorithm. *Journal of Korea Statistical Society*, **17**, 667-678.

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*, 1-94.

Lim, S. Y., Baek, J. S. and Kim, M. S. (2010). Video character recognition improvement by support vector machines and regularized discriminant analysis. *Proceedings of Journal of Korean Data & Information Science Society May 28-29*, **2010**, 1-10.

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.

Schölkopf, B. and Smola, A. J. (2001). *Learning with kernels: Support vector machines, regularization, optimiaztion, 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.

Tan, P., Steinbach, M. and Kumar V. (2006). *Introduction to data mining*, Addison-Wesley.

Vapnik, V. (1998). *Statistical learning theory*, John wiley & sons, New York.

Journal of the Korean Data & Information Science Society
2010, **21**(6), 1271–1280

한국데이터정보과학회지

Support vector machine and multifactor dimensionality reduction for detecting major gene interactions of continuous data

Jea Young Lee[1] · Jong Hyeong Lee[2]

[12]Department of Statistics, Yeungnam University

Received 8 October 2010, revised 18 November 2010, accepted 23 November 2010

Abstract

We have used multifactor dimensionality reduction (MDR) method to study gene-gene interaction effect of statistical model in general. But, MDR method could not be applied in the continuous data. In this paper, continuous-type data by the support vector machine (SVM) algorithm are proposed to the MDR method which provides an introduction to the technique. Also we apply the method on the identify major interaction effects of single nucleotide polymorphisms (SNPs) responsible for economic traits in a Korean cattle population.

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

[1] Professor, Department of Statistics, Yeungnam University, Kyungsan, Korea. E-mail: jlee@yu.ac.kr
[2] Graduate, Department of Statistics, Yeungnam University, Kyungsan, Korea.