Journal of the Korean Data & Information Science Society
2014, **25**(6), 1385–1395

http://dx.doi.org/10.7465/jkdi.2014.25.6.1385
한국데이터정보과학회지

데이터마이닝을 이용한 한우의 우수 지방산합성효소 유전자 조합 선별

김병두[1] · 김현지[2] · 이성원[3] · 이제영[4]

[1]경일대학교 자연계열자율전공학과 · [24]영남대학교 통계학과 · [3]경운대학교 컴퓨터공학과

접수 2014년 7월 18일, 수정 2014년 8월 29일, 게재확정 2014년 10월 21일

요 약

가축의 경제적인 특성은 환경적인 요인과 유전적인 요인의 영향을 받으며, 또한 하나의 유전자가 아닌 여러 유전자의 상호작용의 영향을 받는다고 알려져 있다. 본 논문에서는 선형회귀모형을 활용하여 환경적인 요인을 보정한 자료로 한우의 맛과 육질에 영향을 준다고 밝혀진 지방산합성효소의 단일염기다형성 5개를 이용해 한우의 경제 형질에 영향을 미치는 우수 유전자 조합을 선별하고 우수 유전자형을 밝힌다. 이를 위해 데이터마이닝 기법인 인공신경망, 로지스틱 회귀모형, C5.0, CART 기법을 이용하였다. 공정한 모형 평가를 위해 전체 데이터를 훈련용 데이터 (60%)와 검증용 데이터 (40%)로 나누었고, 훈련용 데이터에서 설정된 모형을 검증용 데이터에 적용시켜 정확도를 비교하였다. 그 결과 C5.0이 최적 모형으로 선정되었으며, C5.0의 의사결정나무를 통해 우수 유전자 조합을 선별하였다.

주요용어: 단일염기다형성, 데이터마이닝, 지방산합성효소, CART.

1. 서론

단일불포화지방산 (monounsaturated fatty acid; MUFA)은 소고기의 맛과 부드러움에 영향을 미치며, 올레인산 (oleic acid; C18:1)은 MUFA의 중심 역할로 요리된 소고기 향의 원인이 된다 (Melton 등, 1982; Mandell 등, 1998; Matsusushi 등, 2011; Oh 등, 2011). 그리고 소고기의 근내지방도 (marbling score; MS)도 소고기의 품질에 주요한 지표가 되고 있다. 이러한 한우의 경제적인 특성은 단일 유전자의 효과가 아니라 여러 유전자의 상호작용으로 일어난다. 따라서 유전자들의 상호작용은 경제형질의 특성을 발견하는데 중요한 역할을 한다.

본 연구에서는 한우의 맛과 육질에 영향을 미치는 유전적인 요인을 찾고자 데이터마이닝 기법 (인공신경망, 로지스틱 회귀모형, C5.0, CART)을 이용하여 경제 형질인 올레인산 (C18:1), 단일불포화지방산 (MUFA), 근내지방도 (MS)와 지방산합성효소 (fatty acid synthase; FASN)의 5가지 단일염기다형성 (single nucleotide polymorphism; SNP)과의 연관성을 알아보고, 각 경제형질에 영향을 미치는 우수한 단일 SNP와 SNP 조합 그리고 우수 유전자형을 선별하였다. 더불어 선별된 우수 유전자형이 경제 형질의 가치를 높인다는 것을 뒷받침하기 위해서, t-검정과 순열검정을 실시하여 통계적으로 유의미한 차이를 갖는지 확인하였다. 본 연구에서 사용한 자료는, 한우의 경제 형질은 유전적인 요인 뿐만아니

[1] (712-701) 경북 경산시 하양읍 가마실길 50, 경일대학교 자연계열자율전공학과, 교수.
[2] (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 석사과정.
[3] (730-739) 경북 구미시 산동면 강동로 730번지, 경운대학교 컴퓨터공학과, 조교수.
[4] 교신저자: (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 교수. E-mail: jlee@yu.ac.kr

라 환경적인 요인들의 영향도 받기 때문에, 환경적인 요인을 배제하고 한우의 경제 형질에 영향을 미치는 유전적인 요인만을 찾기 위해서 선형회귀모형을 활용하여 환경적인 요인인 장소와 일령을 보정한 것을 이용했다 (Lee와 Jin, 2012).

본 연구는 다음과 같이 구성되었다. 2절에서는 우수 유전자 조합 선별을 위해서 사용한 통계적 방법들을 소개하고, 3절에서는 환경적인 요인을 보정한 한우자료에 데이터마이닝 기법들을 적용하고, 각 기법들의 정확도를 비교하여 최종모형을 선택한다. 4절에서는 선택된 모형을 이용해 한우의 경제 형질에 영향을 미치는 우수 유전자조합과 유전자형을 선별한다. 5절에서는 연구의 결과를 요약한다.

2. 우수 유전자 조합 선별을 위한 통계적 방법

2.1절에서는 한우의 환경적인 요인을 보정하기 위해 사용한 선형 회귀모형을 소개하고, 2.2절에서는 모형구축을 위해 사용된 데이터마이닝 기법들을 소개한다. 마지막으로 2.3절에서는 순열검정의 특징과 절차를 소개한다.

2.1. 선형 회귀 모형을 이용한 보정

유전분석에서 어느 개체의 표현형은 그 개체의 유전적인 효과와 환경적인 효과에 의해 결정된다. 즉, 개체의 표현형은 다음과 같이 나타낼 수 있다.

$$P = F + G$$

P는 개체의 표현형 (phenotype)이고 E는 환경적인 효과 (environmental effect), G는 유전적인 효과 (genetic effect)이다. 그래서 한우의 경제형질에 대한 연구에서는 다음과 같은 선형 회귀모형을 사용한다 (Casas 등, 2005).

$$y_{ijk} = \mu + farm_i + \beta \times age + SNP_j + e_{ijk}$$
$$i = 1, \cdots, f,\ j = 1, \cdots, m,\ k = 1, \cdots, n$$

y_{ijk}는 한우의 경제형질이고, $farm_i$는 장소 (17군데)의 고정효과이며 β는 일령 (age)에 대한 회귀계수, SNP_j는 FASN의 5가지 유전적 고정효과, e_{ijk}는 $N(0, \sigma^2)$인 확률변수이다. 여기서 장소와 일령은 환경적인 효과이고, SNP는 유전적인 효과이다. 그러나 우리가 관심을 가지고 있는 부분은 한우의 경제형질에 영향을 주는 유전적인 효과를 밝혀내는 것이기 때문에 환경적인 요인을 보정한 아래의 모형을 연구에 이용한다 (Matsuhashi 등, 2011).

$$y_{ijk} - (farm_i + \beta \times age) = \mu + SNP_j + e_{ijk}$$
$$i = 1, \cdots, f,\ j = 1, \cdots, m,\ k = 1, \cdots, n$$

2.2. 데이터마이닝 기법 소개

인공신경망은 인간의 신경-두뇌 시스템을 흉내 낸 것으로, 몇 개의 뉴런 (neuron)과 이것들이 배열된 층 (layer)으로 구성된다 (Sarle, 1994; Tan 등, 2006; Heo와 Lee, 2008; Park 등, 2011). 각 뉴런은 특정의 작업을 수행하고 신경망은 이들 뉴런을 연결함으로써 자극과 반응간의 관계를 학습하고, 새로운 데이터에 대한 분류·추정 및 예측을 하게 된다. 여기서 분류는 목표변수가 이산형인 경우가 되며, 예측은 목표변수가 연속형인 경우를 일컫는다.

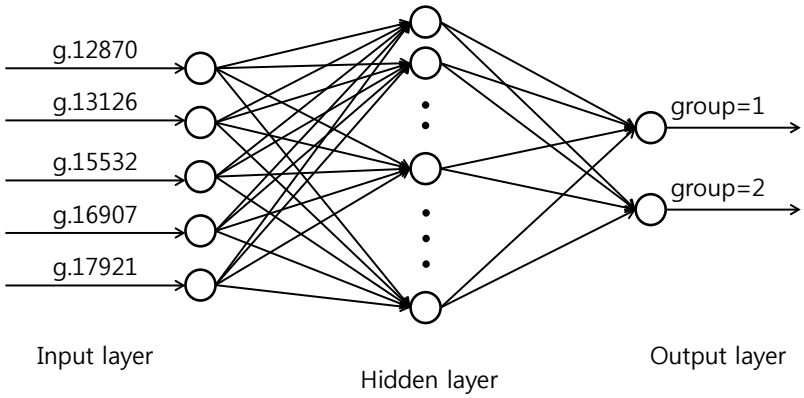

Figure 2.1 Structure of neural network

Figure 2.1은 본 논문의 자료를 신경망으로 표현한 것인데, 신경망 모형은 입력 층 (input layer), 은닉 층 (hidden layer), 출력 층 (output layer) 등 3개 층으로 구성되어 있고 각 층에 몇 개씩의 뉴런이 들어있다. 원으로 표시된 것을 노드라고하며, 입력 층에서 FASN의 5가지 SNP가 5개의 입력노드로 표현되며, 이 입력노드가 어떤 자극을 접수하면 은닉 층의 은닉노드가 각 입력노드로부터 전달되는 신호들을 모아 선형결합을 한다. 그리고 이 신호를 최종적으로 출력 층에서 출력노드가 신호를 전달받아 결합함으로써 최종 반응을 내보내게 된다. 이때 출력노드는 이분형으로 표현된 각 경제형질이다. 신경망은 특히 복잡한 비선형에 적합 시 좋은 결과를 얻을 수 있다.

로지스틱 회귀모형은 임상 연구 자료에서 중요한 요인들을 식별하는 탐색적 분석에 많이 적용된다 (Berson 등, 2000; Lee 등, 2005; Heo와 Lee, 2008). 로지스틱 회귀모형에서 종속변수가 이항자료인 형태가 가장 일반적이며, 본 연구도 이항자료를 이용하므로 이항반응에 대해서만 살펴보기로 한다. 이항반응변수를 종종 베르누이 변수라고도 한다. 이 변수에 대한 분포는 성공에 대한 확률 $P(Y=0|x) = 1-p_x$와 실패에 대한 확률 p_x로 명시된다. 여기서 성공확률 p_x에 대해 $p_x = \alpha + \beta x$와 같은 선형 확률모형을 생각할 수 있다.

하지만 이 모형은 구조적인 결함을 가진다. p_x가 0과 1사이의 값을 가지고 $\alpha + \beta x$는 실수 전체의 값을 가지므로 성공확률 p_x가 범위 외의 값을 가질 수 있게 된다. 따라서 이런 식 대신에 성공확률 p_x에 대해 k개의 설명변수 $x_1, x_2, \cdots, x_k$와 비선형인 아래의 식과 같은 함수를 생각할 수 있다.

$$p_x = \frac{exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$$

위 식에서 성공확률에 대한 오즈(odds)를 구하고, 이 오즈에 로그를 씌우면 아래의 선형 식으로 타나낼 수 있으며, 이 식을 로지스틱 회귀모형이라고 한다.

$$\log \frac{p_x}{1-p_x} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

본 논문에서의 종속변수는 각 경제형질을 두 그룹으로 나눈 이항자료이고, 독립변수는 FASN의 5가지 SNP를 이용한다. 마찬가지로 다음의 CART와 C5.0기법에서도 위 종속변수와 독립변수를 사용한다.

CART (classification and regression tree)는 설명변수들과 목표변수로 이루어진 자료들에서 설명변수들의 특성에 따라 자료들을 이지분류 (binary split)하여, 2개의 하위노드를 생산하는 과정을 반복하여 자료들을 목표변수의 값이 유사한 부분집합으로 만드는 방법이다. CART의 알고리즘은 마디의 순수함을 나타내는 지니계수 (gini index)에 의해 분리여부를 결정한다. 특정 변수에 의해 집단이 구분되면, 구분된 하나의 집단에서 나머지 집단의 개체가 선택될 확률을 계산하여 집단을 분리한다. 집단이 순수할수록 지니계수의 값이 작아지며 확률 또한 작아지게 된다 (Berson 등, 2000). 따라서 CART는 지니계수를 가장 감소시켜 주는 설명변수와 그 변수의 최적분리를 자식노드로 선택한다. 만일 목표변수의 범주가 k개로 분할되고, i번째 범주에 분류될 확률이 p_1, p_2, $\cdots$, p_k일 때, 지니계수는 아래의 식과 같이 정의된다 (Breiman 등, 1984; Heo와 Lee, 2008).

$$G = 1 - \sum_{i=1}^{k} p_i^2$$

C5.0 알고리즘은 정보이론 (information theory)에 따른 엔트로피 (entropy)개념을 이용하여 마디의 정보량에 따른 엔트로피 지수에 의해 분리가 된다. 분리된 마디에서의 집단의 정보량을 많이 가질수록 엔트로피 지수는 작아지게 되며 부족한 것이 없는 완전한 정보를 얻었음을 뜻한다 (Freund와 Mason, 1999; Berson 등, 2000; Quinlan, 1993; Heo와 Lee, 2008).

만일 자료 D가 목표변수에 의하여 범주가 k개로 분할되고, i번째 범주에 분류될 확률이 p_1, p_2, $\cdots$, p_k일 때, 자료 D의 엔트로피 지수는 아래 $info(D)$와 같이 정의된다. 또한 자료 D가 변수 x_i값을 바탕으로 자료를 분할하여 D_1, D_2, $\cdots$, D_n을 얻는다고 했을 때, $|D_j|$를 1개의 분할 자료 D의 크기라고 하면 변수 x_i값을 바탕으로 분할된 $D_1, D_2, \cdots, D_n$의 엔트로피는 아래 $info\ x_i(D)$와 같이 계산된다.

$$info(D) = -\sum_{i=1}^{k} p_i(\log p_i)$$

$$info\ x_i(D) = \sum_{j=1}^{n} \frac{|D_j|}{|D|} info(D_j)$$

일반적으로 변수 x_i값을 바탕으로 자료 D를 분할하여 나온 정보와 단순히 자료 D를 구별하는데 필요한 정보의 차이가 나게 되는데, 이러한 정보의 차이를 정보의 이득 (gain)이라고 한다. 이득의 기준이 이론적으로는 명확하나, 많은 수의 범주를 갖는 예측변수를 선호하므로 실제로 이득기준을 그대로 사용하지는 않으므로, 이득비율을 사용하여 실제적인 변수선택의 기준으로 한다. 따라서 C5.0은 이득비율이 최대화되는 점에서 데이터의 분할을 선택한다.

2.3. 순열검정

선택된 최적 유전자 조합의 통계적 유의성 검정을 위해서 본 연구에서는 t-검정과 순열검정 (Good, 2000)을 통한 p-값을 계산한다. 순열검정의 절차는 다음과 같다.

단계 1. 가설 설정 - 각 방법을 통해 선별된 우수 유전자 조합이 특정치에 영향력이 있다.

단계 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, 특정치에 영향력이 있다고 판단한다.

단계 3. 선별된 유전자형 조합에 대한 통계량 계산 - 각 방법을 통해 선별된 우수 유전자형 조합의 F 측정치를 계산한다.

단계 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.

단계 5. 결론 (p-값 계산) - 각 F 측정치를 내림차순으로 정렬한 후 기존의 F 측정치와 비교하여 p-값을 구한다.

위의 순열검정을 통해서 각각의 우수한 단일 유전자와 유전자 조합에 대해 p-값을 계산하고, 그 결과를 통해 한우의 경제형질에 영향을 미치는 우수 유전자 조합과 우수 유전자형을 규명한다.

3. 적용 및 결과

3.1절에서는 환경적인 요인을 보정한 한우 자료에 대해서 살펴보고, 3.2절에서는 데이터마이닝 기법인 인공신경망, 로지스틱 회귀모형, CART, C5.0을 적용시키고, 각 모형을 비교하여 정확도가 가장 높은 모형을 최종 모형으로 선택한다. 본 논문에서는 분석을 위해 IBM SPSS Modeler 버전 14.1을 사용하였다.

3.1. 실험자료

본 연구는 경북지역에서 자란 18 아비로부터 얻어진 513 두의 한우의 데이터를 사용하였다. 경제 형질은 한우의 맛과 향에 영향을 준다고 알려져 있는 올레인산 (oleic acid; C18:1)과 단일불포화지방산 (monounsaturated fatty acid; MUFA), 한우의 육질에 영향을 준다고 알려진 근내지방도 (marbling score; MS)를 분석에 사용하였다 (Lee 등, 2011). Table 3.1은 환경적인 요인을 보정한 각 경제형질의 평균과 표준편차를 나타낸 것이다.

Table 3.1 Mean and standard deviation of adjusted traits for environmental factors

Economic trait	N	Mean	SD
C18:1	513	44.30	2.12
MUFA	513	53.50	2.34
MS	513	5.43	1.42

C18:1; oleic acid, MUFA; monounsaturated fatty acid, MS; marbling score

본 연구에서는 각 경제형질을 K-평균 알고리즘으로 이분화한 값을 종속변수로 사용하였다. Table 3.2는 이분화한 데이터의 평균과 표준편차를 나타낸 것이다.

Table 3.2 Mean and standard deviation of traits which are divided into two classes

Economic trait	Group	N	Mean	SD
C18:1	1	360	43.28	1.49
	2	153	46.70	1.31
MUFA	1	314	52.08	1.52
	2	199	55.75	1.49
MS	1	241	4.21	0.84
	2	272	6.51	0.83

C18:1; oleic acid, MUFA; monounsaturated fatty acid, MS; marbling score

유전적인 요인은 최근 한우의 맛과 육질에 영향을 준다고 밝혀진 지방산합성효소 (fatty acid synthase; FASN)의 g.12870T>C, g.13126T>C, g.15532C>A, g.16907T>C 그리고 g.17921G>A로 5가지 단일염기다형성 (single nucleotide polymorphism; SNP)을 이용하였다 (Oh 등, 2011)

3.2. 데이터마이닝 기법 적용 결과 및 최종모형 선택

데이터는 공정한 모형평가를 위해서 훈련용 데이터와 검증용 데이터를 각각 60%, 40%로 분할하여 분석하였다. 그리고 인공신경망, 로지스틱 회귀모형, C5.0, CART 순으로 데이터마이닝 기법을 적용하여 모델을 구축하였다. Table 3.3은 각 경제 형질별로 기법에 따른 훈련용 데이터와 검증용 데이터의 정확도를 계산하여 나타낸 것이다. 인공신경망, 로지스틱 회귀모형, CART, C5.0의 훈련용 데이터의 정확도를 살펴보면 올레인산 (C18:1)은 각각 79.11%, 77.85%, 78.16%, 77.22%로 크게 차이가 없고, 단일불포화지방산 (MUFA)도 각각 71.52%, 72.85%, 71.19%, 70.53%로 크게 차이를 보이지 않는다. 근내지방도 (MS)도 각각 63.25%, 66.89%, 67.88%, 66.89%로 크게 차이를 보이지 않는다. 반면에 검증용 데이터의 정확도를 살펴보면 올레인산 (C18:1)은 C5.0이 74.11%로 60%대인 다른 모형과 비교했을때 정확도가 가장 높다. 단일불포화지방산 (MUFA)도 C5.0이 70.62%로 가장 높고, 근내지방도 (MS)도 C5.0이 62.09%로 가장 높다. 그러므로 각 경제 형질을 종합적으로 봤을 때 정확도가 가장 높게 나타난 C5.0 기법을 최종모형으로 선택하여 경제 형질에 영향을 미치는 우수한 단일 SNP와 SNP조합을 선별하였다.

Table 3.3 Comparison of accuracy for each data mining method of each economic trait

Economic trait	Model	Train data accuracy	Test data accuracy
C18:1	Neural network	79.11	69.54
	Logistic regression	77.85	68.02
	CART	78.16	68.02
	C5.0	77.22	74.11
MUFA	Neural network	71.52	69.67
	Logistic regression	72.85	69.19
	CART	71.19	69.67
	C5.0	70.53	70.62
MS	Neural network	63.25	54.98
	Logistic regression	66.89	59.72
	CART	67.88	58.29
	C5.0	66.89	62.09

C18:1; oleic acid, MUFA; monounsaturated fatty acid, MS; marbling score

4. C5.0을 이용한 우수 FASN 유전자형 선별

4절에서는 C5.0 기법을 통하여 각 경제 형질에 영향을 주는 우수한 단일 SNP와 SNP조합을 선별하고, 우수 유전자형을 밝힌다. 그리고 t-검정과 순열검정을 실시하여 우수 유전자형의 통계적인 유의성을 살펴보았다.

4.1. C5.0기법 적용

다음은 한우 데이터에 C5.0 기법을 적용시킨 결과이다. Figure 4.1∼4.3은 각 경제 형질별 의사결정나무이다.

Figure 4.1에서 올레인산 (C18:1)의 의사결정나무를 분석해보면, g.13126T>C가 가장 영향력있는 SNP로 나타났고 g.13126T>C의 유전자형인 TT가 그룹 2로 분류될 확률을 28.5%에서 53.6%까지 높여주어 우수 유전자형으로 선별되었다. 다음으로 (g.13126T>C, g.15532C>A) 조합이 가장 우수한 SNP 조합으로 타나났고 TTCA, TTAA, TCAA 유전자형이 그룹 2로 분류될 확률을 78.1%까지 높여주어 우수 유전자형으로 선별되었다.

Figure 4.2에서 단일불포화지방산 (MUFA)의 의사결정나무를 분석해보면, C18.1과 마찬가지로 g.13126T>C가 가장 영향력있는 SNP로 나타났고 g.13126T>C의 유전자형인 TT가 그룹 2로 분류

될 확률을 38.1%에서 62.1%까지 높여주어 우수 유전자형으로 선별되었다. 다음으로 (g.13126T>C, g.15532C>A) 조합이 가장 우수한 SNP 조합으로 타나났고 TTCC, TTCA, TTAA, TCAA 유전자형이 그룹 2로 분류될 확률을 80.6%까지 높여주어 우수 유전자형으로 선별되었다.

Figure 4.3에서 근내지방도 (MS)의 의사결정나무를 분석해보면, g.15532C>A가 가장 영향력 있는 SNP로 나타났고 g.15532C>A의 유전자형인 CA, AA가 그룹 2로 분류될 확률을 54%에서 80%까지 높여주어 우수 유전자형으로 선별되었다. 다음으로 (g.12870T>C, g.15532C>A) 조합이 가장 우수한 SNP 조합으로 타나났고 CCCC, TCCA, CCCA, TTAA, TCAA, CCAA 유전자형이 우수 유전자형으로 선별되었다.

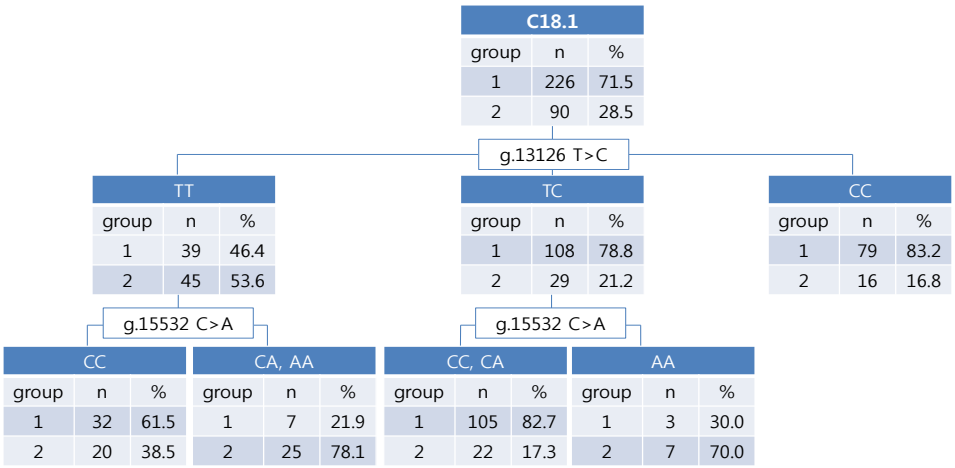

Figure 4.1 Decision tree of oleic acid (C18:1)

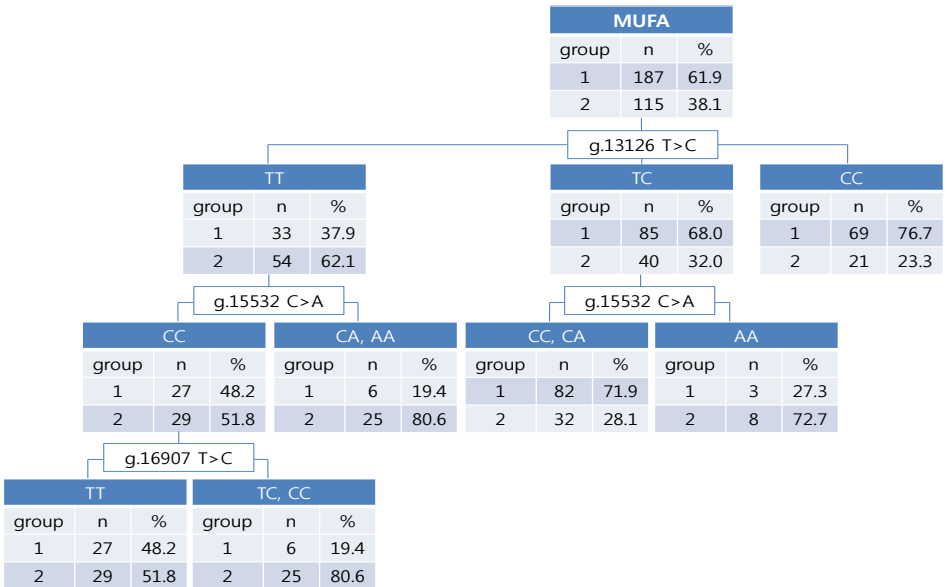

Figure 4.2 Decision tree of monounsaturated fatty acid (MUFA)

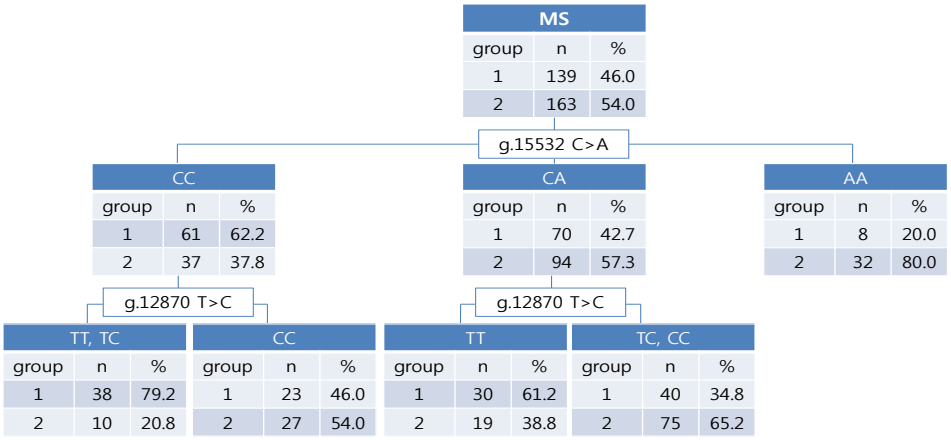

Figure 4.3 Decision tree of marbling score (MS)

4.2. t-검정과 순열검정

4.1절에서 선별된 우수 유전자형이 경제형질의 가치를 높이는지를 확인하기 위해서 *t*-검정과 순열검정을 통해 살펴봤다. table 4.1은 선별된 유전자형이 통계적으로 유의미한지 알아보기 위해 *t*-검정과 순열검정을 실시한 결과이다.

먼저 단일 SNP의 결과를 살펴보면, 올레인산 (C18:1)과 단일불포화지방산 (MUFA)에 영향을 주는 것으로 나타난 g.13126T>C에서는 우수 유전자형인 TT가 각각 평균 45.57, 54.84로 다른 유전자형들에 비해 가치가 유의미하게 높았고 (p <0.001), 근내지방도 (MS)에 영향을 주는 것으로 나타난 g.15532C>A에서는 우수 유전자형인 CA, AA가 평균 5.48로 다른 유전자형들에 비해 가치가 높았다 (p <0.001).

SNP간의 상호작용을 고려한 SNP 조합의 결과를 살펴보면, 올레인산 (C18:1)에 영향을 주는 것으로 나타난 (g.13126T>C, g.15532C>A)조합에서는 우수 유전자형으로 선별된 TTCA, TTAA가 평균 46.49로 다른 유전자형들에 비해 가치가 유의미하게 높았고 (p <0.001), 단일불포화지방산 (MUFA)에 영향을 주는 것으로 나타난 (g.13126T>C, g.15532C>A)조합에서는 우수 유전자형으로 선별된 TTCC, TTCA, TTAA, TCAA가 평균 54.79로 다른 유전자형들에 비해 가치가 높았다 (p<0.001). 근내지방도 (MS)에 영향을 주는 것으로 나타난 (g.12870T>C, g.15532C>A)조합에서는 우수 유전자형으로 선별된 TTAA, TCCA, TCAA, CCCC, CCCA, CCAA가 평균 5.71로 다른 유전자형들에 비해 가치가 높았다 (p <0.001).

Table 4.1 The superior genotype of each economic trait and permutation test

Economic trait	SNP combination	Superior genotypes	N	Mean	SD	*t*-test (*p*-value)	permutation (*p*-value)
C18.1	g.13126T>C	TT	141	45.57	1.88	<0.001	<0.001
		Others	372	43.82	2.02		
MUFA	g.13126T>C	TT	141	54.84	2.05	<0.001	<0.001
		Others	372	53.00	2.24		
MS	g.15532C>A	CA, AA	279	5.48	1.43	<0.001	<0.001
		Others	169	5.10	1.37		
C18.1	g.13126T>C g.15532C>A	TTCA, TTAA	58	46.49	2.10	<0.001	<0.001
		Others	455	44.02	1.96		
MUFA	g.13126T>C g.15532C>A	TTCC, TTCA, TTAA, TCAA	158	54.79	2.01	<0.001	<0.001
		Others	355	52.93	2.25		
MS	g.12870T>C g.15532C>A	TTAA, TCCA, TCAA, CCCC, CCCA, CCAA	338	5.71	1.34	<0.001	<0.001
		Others	175	4.87	1.41		

SNP; single nucleotide polymorphism, C18:1; oleic acid, MUFA; monounsaturated fatty acid, MS; marbling score

5. 결론

본 연구는 한우의 품질에 영향을 미치는 유전적인 요인들을 밝히고자, 한우의 맛과 육질에 영향을 주는 것으로 알려진 올레인산 (C18:1), 단일불포화지방산 (MUFA) 그리고 근내지방도 (MS)에 초점을 맞추어, 이들 경제형질에 영향을 미치는 우수 유전자 조합과 우수 유전자형을 알아보고자 했다. 이때 경제 형질에 영향을 미치는 유전자는 소 염색체 19번에 존재하는 지방산합성효소 (FASN)에서 5가지 단일염기다형성 (SNP)를 이용했다. 그리고 한우의 품질에는 유전적인 요인과 환경적인 요인이 함께 영향을 미치는데, 특히 유전적인 요인만을 고려하기 위해서 선형회귀모형을 통해 환경적인 요인을 보정하여 분석에 사용하였다. 분석은 데이터마이닝 기법 중 검증용 데이터의 모형의 정확도가 가장 높았던 C5.0 기법을 최종 선택하여 사용했고, 의사결정나무를 통해 경제 형질을 가장 분류를 잘한 단일 유전자와 유전자 조합을 선별하고, 세부적으로 우수 유전자형을 찾았다. 그 결과 올레인산 (C18:1)과 단일불포화지방산 (MUFA)에 영향을 주는 우수 단일 유전자는 g.13126T>C가 선별되었고, 경제 형질의 가치를 높이는 우수 유전자형으로는 TT가 선별되었다. 근내지방도 (MS)에 영향을 주는 우수 단일 유전자는 g.15532C>A가 선별되었고, 우수 유전자형으로는 CA와 AA가 선별되었다. 다음으로 유전자 간의 상호작용을 고려하여 선별한 우수 유전자 조합으로는 올레인산 (C18:1)과 단일불포화지방산 (MUFA)에서 (g.13126T>C, g.15532C>A) 조합이 선별되었고, 우수 유전자형으로는 올레인산 (C18:1)은 TTCA, TTAA가 선별되었고, 단일불포화지방산 (MUFA)는 TTCC, TTCA, TTAA, TCAA가 선별되었다. 그리고 근내지방도 (MS)의 우수 유전자 조합으로는 (g.12870T>C, g.15532C>A) 조합이 선별되었고, 우수 유전자형으로는 TTAA, TCCA, TCAA, CCCC, CCCA, CCAA가 선별되었다. 이렇게 선별된 우수 유전자형들이 경제 형질에 유의미한 영향을 미치는지 확인하기 위해서 t-검정과 순열검정을 실시하였고, 그 결과 모든 유전자형의 p-값이 0.001에 근접하여 통계적으로 유의미하게 경제 형질의 가치를 높인다는 것을 확인하였다.

또한 유전자 간의 상호작용을 고려한 우수 유전자 조합이 상호작용을 고려하지 않은 단일 유전자와 비교했을 때, 한우의 경제 형질의 가치를 조금 더 향상시킨다는 것을 확인 할 수 있다. 즉, 한우의 경제 형질의 가치는 단일 유전자 보다는 유전자 간의 상호작용 효과에 의해 더 많은 영향을 받는다는 사실을 알 수 있다.

References

Berson, A., Smith, S. and Thearling, K. (2000). *Building data mining applications for CRM*, McGraw-Hill, New York.

Breiman, L., Friedman, J. H., Olshen, R. and Stone, C. J. (1984). *Classification and regression tree*, Chapman & Hall, New York.

Casas, E., White, S. N., Riley, D. G., Smith, T. P. L., Brenneman, R. A., Olson, T. A., Johnson, D. D., Coleman, S. W., Bennett, G. L. and Chase, C. C. (2005). Assessment of single nucleotide polymorphisms in genes residing on choromosomes 14 and 29 for association with carcass composition traits in Bos indicus cattle. *Journal of Animal Scuence*, **83**, 13-19.

Freund, Y. and Mason, L. (1999). The alternating decision tree learning algorithm. *Proceedings of the Sixteenth International Conference on Machine Learning*, **99**, 121-133.

Good, P. (2000). *Permutation test : A practical guide to resampling methods for testing hypotheses*, Springer-Verlag, New York.

Heo, M. H. and Lee, Y. G. (2008). *Data mining modeling and example*, Hannarae, Seoul.

Lee, J. W., Park, M. R. and Yoo, H. N. (2005). *Statistical methods for life science research*, Free Academy, Seoul.

Lee, J. Y. and Jin, M. H. (2012). Major gene interaction identification in Hanwoo by adjusted environmental effects. *Journal of the Korean Data & Information Science Society*, **23**, 467-474.

Lee, Y. S., Oh, D. Y. and Yeo, J. S. (2011). Study on identification of candidate DNA marker related with beef quality in QTL region of BTA 2 in Hanwoo population. *Journal of the Korean Data & Information Science Society*, **22**, 661-669.

Mandell, I., Buchanan-Smith, G. and C. P. Campbell. 1998. Effects of forage vs grain feeding on carcass characteristics, fatty acid composition, and beef quality in Limousin-cross steers when time on feed is controlled. *Journal of Animal Science*, **76**, 2619-2630.

Matsuhashi. T., Maruyama. S., Uemoto. Y., Kobayashi. N., Mannen. H., Abe. T., Sakaguchi. S. and Kobayashi. E. (2011). Effects of bovine fatty acid synthase, stearoyl-coenzyme A desaturase, sterol regulatory element-binding protein 1, and growth ghormone gene polymorphisms on fatty acid composition and carcass traits in Japanese Black cattle. *Journal of Animal Science*, **89**, 12-22.

Melton, S. L., Amiri, M., Davis, G. W. and Backus, W. R. (1982). Flavor and chemical characteristics of ground beef from grass-, forage-grain- and grain-finished steers. *Journal of Animal Science*, **55**, 77-87.

Oh, D. Y., Lee, Y. S., La, B. M., Yeo, J. S., Chung, E. Y., Kim, Y. Y. and Lee, C. Y. (2011). Fatty acid composition of beef is associated with exonic nucleotide variants of the gene encoding FASN. *Molecular Biology Reports*, **39**, 4083-4090.

Park, I. S., Han, J. T., Sohn, H. S. and Kang, S. B. (2011). Developing the administrative model using the data mining technique for injury in National Health Insurance. *Journal of the Korean Data & Information Science Society*, **23**, 467-476.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*, Morgan-Kaufmann Publishers, San Mateo, CA.

Sarle, W. S. (1994). Neural networks and statistical models. *Proceedings of the 19th Annual SAS Users Group International Conference*, 1-13.

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

Journal of the Korean Data & Information Science Society
2014, **25**(6), 1385–1395

http://dx.doi.org/10.7465/jkdi.2014.25.6.1385
한국데이터정보과학회지

Major gene identification for FASN gene in Korean cattles by data mining

Byung-Doo Kim[1] · Hyun-Ji Kim[2] · Seong-Won Lee[3] · Jea-Young Lee[4]

[1]Department of liberal arts in engineering, Kyungil University
[24]Department of Statistics, Yeungnam University
[3]Department of Computer Engineering, Kyungwoon University

Received 18 July 2014, revised 29 August 2014, accepted 21 October 2014

Abstract

Economic traits of livestock are affected by environmental factors and genetic factors. In addition, it is not affected by one gene, but is affected by interaction of genes. We used a linear regression model in order to adjust environmental factors. And, in order to identify gene-gene interaction effect, we applied data mining techniques such as neural network, logistic regression, CART and C5.0 using five-SNPs (single nucleotide polymorphism) of FASN (fatty acid synthase). We divided total data into training (60%) and testing (40%) data, and applied the model which was designed by training data to testing data. By the comparison of prediction accuracy, C5.0 was identified as the best model. It were selected superior genotype using the decision tree.

Keywords: CART, data mining, fatty acid synthase, single nucleotide polymorphism.

[1] Professor, Department of liberal arts in engineering, Kyungil University, Kyungsan, 712-701, Korea.
[2] Graduate student, Department of Statistics, Yeungnam University, Kyungsan 712-749, Korea.
[3] Assistant professor, Department of Computer Engineering, Kyungwoon University, Gumi 730-739, Korea.
[4] Corresponding author: Professor, Department of Statistics, Yeungnam University, Kyungsan 712-749, Korea. E-mail: jlee@yu.ac.kr