Journal of the Korean Data & Information Science Society
2009, **20**(4), 639–647

한국데이터정보과학회지

유전자 알고리즘에 의한 우수 유전자형 선별

이제영[1] · 고진영[2]

영남대학교 통계학과

접수 2009년 4월 2일, 수정 2009년 7월 1일, 게재확정 2009년 7월 7일

요 약

컴퓨터공학의 발전으로 인해, 여러 개의 변수가 존재하는 비선형 문제와 같은 최적해 탐색과 최적화에 사용되는 유전자 알고리즘은 많은 분야에서 활발하게 응용되고 있다. 그 중, 데이터마이닝분야에서 유전자 알고리즘을 이용하여 정확도를 최대로 하는 입력변수 선택방법과 여러 예측모형을 통합하는 방법 등이 제시되었다. 한편, 우리나라 축산업을 대표하는 한우의 유전자원 보존과 능력향상을 위해서는 다음세대에 유전이 되는 단일염기다형성에서 특정 유전자형을 가진 한우가 경제형질이 우수한지를 찾아낼 필요가 있다. 이에 따라, 유전자 알고리즘을 이용하여 한우의 경제형질에 가장 많은 영향을 주는 단일염기다형성 조합마커의 유전자형을 선택하는 방법을 제시하였다. 그리고 실제 한우 유전 데이터에 적용하여 주요 단일염기다형성 조합마커에서 우수 유전자형들을 선별하였다.

주요용어: 단일염기다형성, 유전자 알고리즘, 유전형질, 의사결정나무분석.

1. 서 론

컴퓨터공학의 발전으로 인해, 수 만년동안 걸쳐 인간이나 생물이 자연 진화하거나 도태되는 모든 일을 소프트웨어 환경에서 불과 몇 시간 또는 몇 분 안에 흉내 내어 최적의 해를 찾아내도록 설계된 것이 유전자 알고리즘이다. 이는 1950년대와 1960년대에 이루어진 진화 연산의 연구에서부터 시작되었다. 진화가 문제해결을 위해 최적화 도구로 사용될 수 있다는 아이디어로 자연의 유전적 변이와 자연선택에서부터 만들어진 연산자들을 사용함으로 주어진 문제에 대한 잠재 해의 집합을 진화시켜 나가는 방법이다.

최초의 유전자 알고리즘 (genetic algorithm)은 자연생태계에서 관찰된 진화 방법 및 유전학 이론을 컴퓨터 이용이 가능하도록 최적해 탐색 문제에 도입한 것으로서 Holland (1975)에 의해서 제안되었다. 이러한 유전자 알고리즘에서는 기본적으로 단일해가 아닌 잠재적인 해집합을 이용하여 선택 (selection), 교배 (crossover) 및 돌연변이 (mutation) 등의 유전 연산을 통해 변환하여, 다음세대의 개체군이 생산되는 절차를 반복하며 최적의 해를 찾아낸다.

수학적으로 해결하기 힘든 문제에 관한 최적해의 탐색과 일정예산에서 물품의 구매와 같은 최적화에 사용되는 유전자 알고리즘은 신경망 이론이나 퍼지이론 등과 마찬가지로 학문의 거의 전 분야에서 활발하게 응용되고 있다. 예를 들어 가스파이프 라인 제어의 최적화 (Goldberg, 1983), 순환방문 판매원 문제 (Goldberg과 Lingle, 1985; Whitely 등, 1989) 등이 있다. 또한 개념과 이론이 비교적 단순하고, 해의 탐색 성능이 우수하여 광범위한 검색 범위를 가진 어플리케이션에서 효과적으로 해를 찾는데 매우 유용한 기법임이 많은 선행 연구들에 의해 증명된 바 있다 (Colin, 1994; Han 등, 1997; Koze, 1993; Shin과 Han, 1998). 그리고 데이터마이닝 분야에서 예측율은 높지만 국소탐색으로 국지적 최적화 (local

[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, 대학원, 석사.

optima)에 빠질 수 있는 단점을 가진 인공신경망 모형과 상호 보완하고자 하는 연구가 있었다 (Kitano, 1992; Barbro 등, 1996). 그리고 기업부도 예측 모형을 중심으로 인공신경망 기법의 최적 변수 조합을 선정하기 위하여 유전자 알고리즘을 이용하는 방법론을 제시하였고 (홍승현과 신경식, 2003), 데이터 불균형 문제 해소기법들을 유전자 알고리즘을 통해 최적의 결합비율로 적용하여 소수 범주에 대한 예측 정확성을 높이는 방법을 제시하였다 (장영식 등, 2008).

또한 생명공학의 발전으로 유전정보를 밝혀내려는 지놈 (genome) 연구가 활발히 이루어지고 있는데, 몇몇 나라에서는 인간 지놈 프로젝트 (human genome progect) 뿐만 아니라 경제성이 높은 동물의 지놈 프로젝트도 추진되었고 많은 연구가 이루어져왔다. 우리나라에서도 가장 경쟁력이 확보되고 경제성이 높은 것으로 확인된 한우를 대상으로 유전자지도 작성 (gene mapping)이 시도되어 왔다 (Kim 등, 2000; Yeo 등, 2004). 그리고 많은 세대를 거듭할수록 대립유전자의 유전이 안정적으로 발생되어지고 개체의 기능적인 유전적 가치를 직접적으로 추정할 수 있는 단일염기다형성 (Single Nucleotide Polymorphism; SNP)을 이용하여 한우의 경제적 특성에 많은 영향을 주는 SNP 조합마커를 찾는 연구가 이루어졌다 (Lee와 Choi, 2007; Lee와 Kim, 2008).

한편, 한우의 품질을 향상하고 우수형질의 한우를 보존하기 위해서는 다음세대에 유전이 되는 SNP에서 특정 유전자형 (genotype)을 가진 한우가 경제형질이 우수한지를 찾아낼 필요가 있다. Lee 등 (2009)에 의해 단일 SNP보다 복합 SNP의 상호작용이 한우의 경제형질에 많은 영향을 준다는 것을 밝혀냈다. 그래서 복합 SNP에 해당하는 유전자형 중에서 가장 한우의 경제형질에 영향을 많이 주는 우수 유전자형을 찾아야 한다. 이를 위해, 본 연구에서는 한우의 경제형질에 영향을 주는 우수 유전자형을 찾기 위해 유전자 알고리즘을 사용하는 방법을 제시한다. 2장에서는 유전자 알고리즘에 대해 알아보고, 이를 한우데이터에 적용하기 위한 방법을 제시한다. 그리고 3장에서는 Kim 등 (2003)에 규명 되어진 한우 염색체 6번에 위치한 후보 QTL (Quantitative Trait Loci)인 ILSTS035와 같은 거리에 있는 SNP들 중 다형성이 나타난 SNP (19_1,18_4,28_2) (Lee 등, 2008)를 2장에서 제시한 방법에 적용하여 우수 유전자형을 선별하였다.

2. 유전자 알고리즘과 우수 유전자형 선별 방법

2.1. 유전자 알고리즘

유전자 알고리즘 (genetic algorithm)은 확률적 탐색이나 학습 및 최적화를 위한 기법 중 한 가지로써, 자연에 잘 적응하는 개체는 생존하고, 그렇지 못하면 도태된다는 찰스 다윈의 적자생존 (survival of the fittest)의 이론과 자손의 형질은 두 부모로부터 받는 유전자에서 유전된다는 멘델의 유전 법칙을 바탕으로 하고 있다 (Goldberg, 1989). 이러한 유전자 알고리즘은 고등생물이 염색체 (chromosome) 내의 유전인자가 교배 (crossover) 및 돌연변이 (mutation)를 이용, 세대 (generation)를 거듭함에 따라 최적 상태로 진화해 나가는 데서 힌트를 얻어 이를 견고한 최적해 탐색에 이용하려는 노력의 일환으로 탄생하게 되었다. 유전자 알고리즘은 해 공간에서 단일 해가 아닌 해집합을 이용하기 때문에 전역적 해 (global optimization)의 발견을 가능케 하고, 최적화 함수 정보를 필요치 않으므로 성능지표 또는 평가함수를 설계자의 의도에 부합하도록 용이하게 정의할 수 있는 장점을 가지고 있다 (최규석 등, 2008).

최적의 해를 탐색하기 위하여, 해집합에서 랜덤하게 추출한 해들로 이루어진 다수의 개체를 생성한다. 이를 초기 개체군 (initial population)이라 하며, 개체군을 선택, 교배 및 돌연변이를 통해 반복적으로 해를 탐색한다. 우수 개체를 선택하기 위해서 각각의 개체들이 찾고자 하는 해와 얼마나 동일한지 평가함수를 통해 평가한다. 선택방법에는 룰렛 선택법, 기대치 선택법, 순위 선택법, 토너먼트 선택법 등이 있다. 그 중 토너먼트 선택법은 개체 집단 중에서 결정된 수 (보통 2)만큼의 개체를 무작위로 선택한 다음, 그들 중에서 가장 높은 평가 값을 가지는 개체를 선택하는 방법이로 초기 개체군과 동일한 개수의

개체군을 재구성한다. 재구성된 개체군을 교배와 돌연변이 등의 유전연산을 통해 이전세대의 개체군과 다른 개체군으로 변환 시킨다.

교배는 미리 설정된 교배율에 따라 재구성된 개체군에서 임의로 선택된 두 개체를 부분적으로 서로 결합하여 새로운 개체를 만드는 과정으로 교배위치에 따라 여러 가지 방법이 있다. 교배율이 너무 크면 새로운 개체가 개체군에 빨리 나타나고 선택연산자가 개선시키는 것보다 더 빨리 성능이 좋은 개체가 무시된다. 반면에 작으면 탐색이 부진하게 될 수 있다. 그러므로 적절한 값으로 설정해 주어야 한다.

돌연변이는 설정된 돌연변이율에 따라 개체내의 임의의 부분을 선택하여 변환시키는 것을 말한다. 이때 확률 값이 너무 작으면 개체군내에서 유일한 값으로 수렴하고, 너무 크면 임의적인 탐색이 되므로 적절하게 설정해야 된다. 이와 같은 절차를 반복해 가면서 문제의 해를 탐색하고 가장 적합한 값을 구할 수 있다.

이런 유전자 알고리즘을 통하여 최적의 해를 찾는 연구에서 해당 문제에 대한 최적의 해를 어떻게 개체로 표현할 것인지와 각 개체가 구하고자 하는 최적의 해에 얼마나 적합한지를 측정할 수 있는 기준, 즉 평가함수에 대한 정의가 무엇보다 중요한 문제이다. 그래서 본 논문에서는 잠재적인 해집합을 한우의 경제형질을 예측하는데 사용되는 복합 SNP의 상호작용변수들로 정의하였다. 하나의 개체는 전체 독립변수들 중에서 임의로 선택된 변수들의 조합으로 이루어지며, 각 개체의 평가를 위해 의사결정나무분석 (C4.5)을 사용하였다. 즉, 개체 내에 선택된 독립변수들을 입력변수로 하여 train데이터로 의사결정나무모형을 생성하고, 평가용자료를 모형에 적용한 정확도에 따라 개체를 평가한다. 최종적으로 선택된 변수로 한우의 경제형질을 예측하는데 사용할 모형이 의사결정나무분석이기 때문에 좀 더 나은 예측력을 얻기 위해 유전자 알고리즘의 평가함수로 의사결정나무분석을 사용하였다. 지금까지 설명한 유전자 알고리즘에 대한 절차를 도식화하면 다음 그림 2.1과 같다.

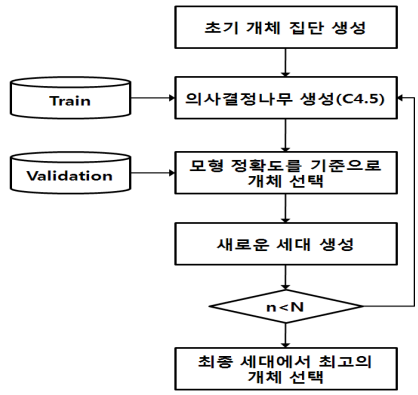

그림 2.1 유전자 알고리즘 절차

유전자 알고리즘에는 미리 정해줘야 하는 매개변수 값들이 존재하는데, 이 매개변수 값들이 결과에 많은 영향을 끼칠 수 있다. 그러나 이론적으로 정해진 값이 없고 실험을 통하여 문제에 적합한 값들을 결정해야 하는 기술적인 문제로 남아 있다 (김여근 등, 1999). 그래서 본 연구에서 유전자 알고리즘의 매개변수 값은 다음 표 2.1과 같이 지정하였다.

이러한 유전자 알고리즘을 이용하여 다음 절에서는 우수 유전자형을 선별하는 방법에 대해 살펴보도록 하자.

표 2.1 유전자 알고리즘 사용을 위한 매개변수

매개 변수	설정 값
개체군의 크기(population size)	20
진화 횟수(max generation)	20
교배점(cross point)	단일점 교배
교배율(cross rate)	0.6
돌연변이율(mutation rate)	0.033
적합도 함수(fitness function)	C4.5 모형 정확도

2.2. 유전자 알고리즘을 이용한 우수 유전자형 선별방법

데이터를 가지고 분석에 앞서 복합 SNPs의 유전자형에 해당하는 변수를 생성한다. 하나의 SNP는 n개의 유전자형을 가지는 범주형 변수로 복합 SNP 조합마커가 가지는 n2개의 유전자형들을 genotype_1 ∼ genotype_n2이라는 새로운 변수를 만든다. 그리고 각 변수의 값으로는 특정 유전자형을 가지면 1, 아니면 0을 가진다. 총 n2개의 유전자형을 입력변수로 하고 한우의 경제형질에 해당하는 1개의 목표변수를 가지는 데이터를 이용하여 분석을 실시한다. 유전자 알고리즘을 이용한 우수 유전자형 선택방법이 결합된 모형의 절차는 다음과 같다. 그리고 제시된 방법을 도식화하면 다음 그림 2.2와 같다.

- 단계 1: 목표변수가 연속형인 경우, 경제형질 변수를 k-means기법을 통해 k가 2인 그룹으로 군집화하고, 이 때 그룹 평균이 높은 그룹을 high, 낮은 그룹은 low로 하는 이분형 변수로 만들어 준다.
- 단계 2: 모형의 정확도를 향상하기 위하여 데이터를 bootstrap방법으로 데이터를10배 boosting하고, 전체데이터를 train 35%, validation 35%, test 30%로 나눈다.
- 단계 3: 유전자 알고리즘 기법을 이용하여 train데이터로 평가함수인 C4.5모형을 만들고, validation데이터를 모형에 적용하여 정확도를 구한다. 최대반복 수 만큼 반복하여 정확도를 최대로 하는 유전자형을 선택한다.
- 단계 4: 선택된 유전자형으로 최종 C4.5모형을 생성한다.
- 단계 5: test데이터를 적용한 모형의 정확도가 가장 높은 SNPs 조합마커에서 선택된 유전자형을 우수 유전자형으로 규명한다.

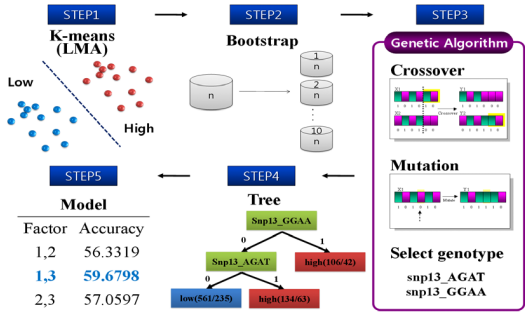

그림 2.2 유전자 알고리즘을 이용한 우수 유전자형 선택방법

유전자 알고리즘을 통해 각각의 SNPs 조합마커에서 우수한 유전자형을 선별하고 모형의 정확도가 가장 높은 SNPs 조합마커를 한우의 경제형질에 가장 큰 영향을 주는 SNPs 조합마커라고 규명할 수 있

다. 이를 검정하기 위해 의사결정 나무분석 모형을 통해 두 그룹으로 나누어 유의성 검정을 하였다. 검정 방법으로 한우의 경제형질에 우수한 영향을 주는 genotype들은 'high' 그룹으로, 나머지 genotype은 'low' 그룹으로 지정하여 Permutation test (Good, 2000)와 스튜던트 t검정을 실시하였다. 그 중에서 Permutation test 절차에 대해 살펴보면 다음과 같다.

- 절차 1. 가설 설정- '선택된 유전자형이 경제형질에 영향력이 있다'를 대립가설로 설정한다.
- 절차 2. 통계량과 기각역 설정- 사용할 통계량은 'high'그룹의 평균을 사용하였으며, 선택된 유전자형에 의한 'high'그룹의 평균이 그룹 간에 데이터를 서로 맞바꾸었을 때 보다 높으면 경제형질에 영향력이 있다고 판단한다.
- 절차 3. 기존 관측치의 통계량 계산- 각 경제형질 데이터 (test데이터)를 의사결정나무분석 모형에 의해 'high'그룹과 'low'그룹으로 나누어 'high'그룹의 평균을 계산한다.
- 절차 4. 데이터의 재배열과 재배열 후의 유의확률 계산- 두 그룹의 데이터를 n개만큼 랜덤 추출하여 그룹을 상호 변경한 후 'high'그룹의 평균을 구한다. 이 과정을 10,000번 반복한다. 10,000개의 평균을 내림차순으로 정렬한 후 기존의 평균과 비교하여 Monte Carlo의 유의확률 값을 구한다.
- 절차 5. 결론- 유의수준 0.05에서 유의확률 값이 작으면 '선택된 유전자형이 경제형질에 영향력이 있다'는 가설을 채택한다.

최종적으로 가장 높은 정확도를 가지는 SNPs 조합마커를 검정을 통해 유의성을 확인하고, 유의성이 확인된 SNPs 조합마커에서 선택된 유전자형을 목표변수에 가장 많은 영향을 주는 우수 유전자형으로 규명한다.

본 논문에서 제시하는 방법을 구현하기 위해 JAVA를 기반으로 한 WEKA (2008)를 사용하였다 WEKA는 Waikato Environment for Analysis의 약어로, 와이카토 대학에서 개발하여 공개된 자바 기반의 데이터마이닝 프로그램이다. 그리고 아래의 그림 2.3은 WEKA프로그램 내의 분석 흐름도를 나타내고 있다

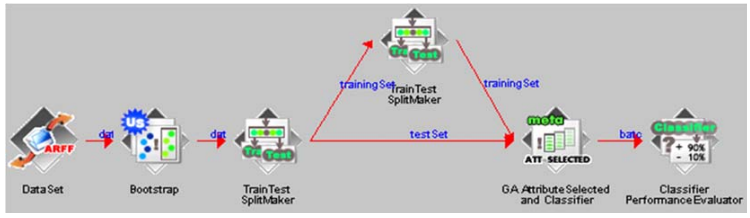

그림 2.3 WEKA프로그램 내의 분석 흐름도

2장에서는 유전자 알고리즘을 소개하고, 유전자 알고리즘과 의사결정나무분석을 이용하여 우수 유전자형을 찾는 방법을 제시하였다. 3장에서는 제시된 방법을 실제 데이터에 적용한 결과를 살펴보도록 한다.

3. 유전자 알고리즘을 이용한 우수 유전자형 선별 및 결과

3.1. 자료

본 연구 데이터는 농협중앙회 가축개량 사무소에서 개발되었고 16 grand-sire half-sibs families로부

터 229두의 수송아지로 구성되었다. 한우의 여러 경제형질인 등심단면적 (LMA: longissimus muscle dorsi area), 도체중 (CWT: carcass cold weight), 일당증체량 (ADG: average daily gain)은 모든 F1자손으로부터 수집되었고 한국축산물등급판정소의 규격에 따라 측정되었다.

현재까지 소에서는 도체형질 (도체중량, 등지방두께, 등심단면적, 일당증체량, 근내지방도)과 연관이 있는 SNPs 조합마커들의 일반 가축에서 평가되거나 적용되고 있다 (Barendse 등, 2004; Page 등, 2004). 따라서 본 연구에서는 EST-based SNP 연관지도 (Snelling 등, 2005)에서 Kim 등 (2003)에 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 (Lee 등, 2008).

3.2. 우수 유전자형 선별

앞에서 제시한 방법을 토대로 각각의 SNPs 조합마커에서 유전자 알고리즘을 통해 한우의 경제형질을 예측하는데 우수한 유전자형을 선별하였다. 우선 LMA를 예측하는데 우수 유전자형으로 SNP (19_1)*SNP (18_4)조합마커에서는 AGTT, GGTT가 선별하였고, SNP (19_1)*SNP (28_2)조합마커에서는 AGAT, GGAA이 선별하였다. 그리고 SNP (18_4)*SNP (28_2)조합마커에서는 CCAA, TTAT이 우수 유전자형으로 선택되었다. 또한 CWT, ADG에서도 표 3.1과 같이 우수 유전자형이 선별되었다.

표 3.1 유전자 알고리즘으로 선택된 우수 유전자형

경제형질	**LMA**	**CWT**	**ADG**
SNP(19_1)*SNP(18_4)	AGTT, GGTT	AACC, AATT, AGTT, GGTT, GGCC	.
SNP(19_1)*SNP(28_2)	**AGAT , GGAA**	**AAAA, AAAT, AGAT, GGAA, GGAT**	**GGAA**
SNP(18_4)*SNP(28_2)	CCAA, TTAT	CCAA, CCAT, TTAT	.

그리고 위의 표3.1에서 선별된 유전자형으로 의사결정나무모형을 만들고 test데이터를 적용한 모형정확도를 구해보았다. 아래의 표 3.2와 같이 각각의 경제형질에서 SNPs 조합마커들의 모형정확도를 비교해 보면 SNP (19_1)*SNP (28_2)조합마커의 정확도가 LMA, CWT, ADG에서 59.6798, 59.9709, 59.8253로 다른 SNPs 조합마커보다 가장 높게 나타났다.

표 3.2 의사결정나무분석의 모형 정확도

	Accuracy		
SNPs 조합마커	LMA	CWT	ADG
SNP(19_1)*SNP(18_4)	56.3319	53.1295	58.9520
SNP(19_1)*SNP(28_2)	**59.6798**	**59.9709**	**59.8253**
SNP(18_4)*SNP(28_2)	57.0597	58.3697	58.9520

따라서, SNP (19_1)*SNP (28_2)조합마커를 한우의 경제형질에 가장 많은 영향을 주는 SNP 조합마커로 규명하였다. 이를 검증하기 위하여 비모수적인 방법으로 permutation검정과 모수적인 방법으로 스튜던트 t검정을 시행하였다. 우선 검증을 위해 의사결정나무모형에서 'high'로 예측된 유전자형들을 'high'그룹으로 정하고 나머지 유전자형들을 'low'그룹으로 분류하였다. 그리고 분류된 두 그룹을 2.2장에서 설명한 permutation절차에 따라 검정을 시행한 결과를 보면 표 3.3과 같다. LMA와 CWT에서 0.0001로 유의수준 0.01에서 매우 유의한 차이가 있다고 나타났다. ADG에서는 0.0051로 유의한 차이가 있다고 나타났다. 그리고 스튜던트 t검정 결과에서도 LMA, CWT에서 두 그룹의 차이는 0.0001, 0.0001로 유의수준 0.01에서 매우 유의한 차이가 있다고 나타났으며, ADG는 0.0160으로 유의한 차이가 있다고 나타났다.

표 3.3 Permutation 검정과 스튜던트 t검정 결과

		LMA	CWT	ADG
permutation	p-value	0.0001	0.0001	0.0047
스튜던트	t value	3.84	4.52	2.41
t검정	(p-value)	(0.0001)	(0.0001)	(0.0160)

따라서 한우의 경제형질인 LMA, CWT, ADG에 가장 많은 영향을 주는 SNPs 조합마커로 SNP (19_1)*SNP (28_2)조합마커를 규명하였고, 이에 포함되는 LMA에 AGAT, GGAA, CWT에 AATT, AGAA, AGTT, GGTT 그리고 ADG에 GGAA를 한우의 경제형질에 가장 많은 영향을 주는 우수 유전자형으로 선별할 수 있다.

4. 결론 및 토의

한우의 품질을 향상하고 우수형질의 한우를 보존하기 위해서 다음세대에 유전이 되는 SNP에서 특정 유전자형을 규명하는 연구를 하였다. 이를 위해, 최적해의 탐색 알고리즘인 유전자 알고리즘과 데이터마이닝의 분류모형인 의사결정나무분석을 결합적으로 활용하여 한우의 경제형질에 우수한 유전자형을 규명하는 방법을 제시하고 이를 실제 한우유전 데이터에 적용해 보았다. 유전자 알고리즘으로 각각의 SNPs 조합마커에서 한우의 경제형질에 영향을 주는 우수 유전자형들을 선별하였고, 의사결정나무 모형의 정확도가 가장 높게 나타난 SNP (19_1)*SNP (28_2)조합마커를 주용 SNPs 조합마커로 규명하였다. 이를 검정하기 두 그룹으로 나누어 permutation 검정과 스튜던트 t검정을 시행하여 매우 유의함을 밝혔다. 따라서, 한우의 경제형질에 가장 많은 영향을 주는 우수 유전자형으로는 SNP (19_1)*SNP (28_2)조합마커에 해당하는 유전자형으로 규명할 수 있으므로, LMA에서 AGAT, GGAA, CWT에서 AATT, AGAA, AGTT, GGTT 그리고 ADG에서 GGAA가 한우의 경제형질에 가장 우수한 영향을 주는 유전자형으로 규명되어졌다. 그러므로 LMA에서 AGAT, GGAA, CWT에서 AATT, AGAA, AGTT, GGTT 그리고 ADG에서 GGAA를 가진 한우를 육성, 배양함으로 최고 품질의 한우를 보존할 수 있을 것이다.

추후 연구로는 본 연구에서 사용한 분석방법을 응용하여 유전자의 상호작용을 분석해 보고, 유전자의 상호작용을 분석방법인 MDR (Multifactor Dimensionarity Reduction), RPM (Restricted Partition Method) 등을 통해 나온 결과와 비교하여 볼 수 있을 것이다.

참고문헌

김여근, 윤복식, 이상복 (1999). <메타 휴리스틱>, 영지문화사, 서울.

장영식, 김종우, 허 준 (2008). 유전자 알고리즘을 활용한 데이터 불균형 해소 기법의 조합적 활용. <지능정보연구>, **14**, 133-154.

최규석, 박종진 (2008). <인공지능시스템>, 21세기사, 파주.

홍승현, 신경식 (2003). 유전자 알고리즘을 활용한 인공신경망 모형 최적 입력변수의 선정. <한국지능정보시스템학회 논문지>, **9**, 227-249.

Barbro, B., Teija, L. and Kaisa, S. (1996). Neural networks and genetic algorithms for bankruptcy predictions. *Proceedings of The third World Congress on Expert Systems*, 123-130.

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.

Colin, A. M. (1994). Genetic algorithms for financial modeling. In Dedoeck, G.J. (Edition), *Trading on the edge*, John Wiley, New York, 148-173.

Goldberg, D. (1983). *Computer-aided gas pipeline operation using genetic algorithms and rule learning*, Ph.D. Thesis, University of Michigan.

Goldberg, D. and Lingle, R. (1985). Alleles, loci, and the travelling salesman problem. *Proceedings of ICGA*, **85**, 154-159.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*, Addison-Wesley, Massachusetts.

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

Han, I., Jo, H. and Shin, K. S. (1997). The hybrid systems for credit rating. *Journal of the Korean Operations Research and Management Science Society*, **22**, 163-173.

Holland, J. H. (1997). *Adaptation in natural and artificial systems*, The University of Michigan Press, Michigan.

Kim, J. W., Jang, T. K., Park, Y. A. and Yeo, J. S. (2000). Linkage mapping of chromosome 6 in the Korean Cattle (Hanwoo). *Asian-Australasian Journal of Animal Sciences*, **13**, 235.

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.

Kitano, H. (1992). *Neurogenetic learning: An integrated method of designing and training neural networks using genetic algorithmes*, TechnicalReport, Carnegie Mellon University.

Koza, J. (1993). *Genetic programming*, The MIT Press, Cambridge.

Lee, J. Y. and Choi, Y. M. (2007). Hanwoo individual identification with DNA marker information. *Journal of the Korean Data and Information Science Society*, **18**, 599-608.

Lee, Y. S., Lee, J. H., Lee, J. Y., Kim, J. J., Park, H. S. and Yeo, J. S. (2008). Identification of Candidate SNP (Single Nucleotide Polymorphism) for Growth and Carcass Traits Related to QTL on Chromosome 6 in Hanwoo (Korean Cattle). *Asian-Australasian Journal of Animal Sciences*, **21**, 1703-1709.

Lee, J. Y., Kim, D. C. (2009). Restricted partition method and gene-gene interaction analysis with Hanwoo economic traits. *Journal of the Korean Data and Information Science Society*, **20**, 171-178.

Page, B. T., Vasas, E., Quaas, R. L., Thallman, R. M., Wheeler, T. L., S- hackelford, 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 infuential industry sires. *Journal of Animal Science*, **82**, 3474-3481.

Shin, K. S. and Han, I. (1998). Using genetic algorithm to support case-based reasoning: Application to corporate bond rating integration. *Proceedings of Second Asia Pacific Decision Sciences Institute (DSI) Conference*, Taipei, 1-11.

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

Whitely, D., Starkweather, T. and Fuquay, D. (1989). Scheduling problems and traveling salesman: The genetic edge recombination operator. *Proceedings of ICGA*, 89.

Yeo, J. S., Lee, J. Y. and Kim, J. W. (1989). DNA marker mining of ILSTS035 microsatellite locus on chromosome 6 of hanwoo cattle. *Journal of Genetics*, **83**, 245-250.

Journal of the Korean Data & Information Science Society
2009, **20**(4), 639–647

한국데이터정보과학회지

Selection of the principal genotype with genetic algorithm

Jea-Young Lee[1] · Jin Young Goh[2]

Department of Statistics, Yeungnam University

Received 2 April 2009, revised 1 July 2009, accepted 7 July 2009

Abstract

From development of computer science, genetic algorithm has been applied to many fields for search like non-linear problem based on various variables and optimization process. Among others, in the data mining field, there are methods to select the best input variables for model accuracy and various predict models which were merged by using the genetic algorithm. In the meantime, to improve and preserve quality of the Hanwoo (Korean cattle) which is represented the agricultural industry in our country, we need to find out outstanding economical traits of Hanwoo in having specific genotype of single nucleotide polymorphism (SNP) which is inherited to next generation. According to, This research proposed the selecting method to find genotype of SNPs marker which affects economical traits of the Hanwoo by using the genetic algorithm. And we selected the best genotypes of the principal SNPs marker by applying to real data on Hanwoo genetic.

Keywords: Decision tree, economic trait, genetic algorithm, genotype, SNP (single nucleotide polymorphism).

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