Korean J Leg Med 2017;41:32-40
https://doi.org/10.7580/kjlm.2017.41.2.32

종 설

데이터마이닝 방법을 이용한 아시아 민족 분류 모형 구축

김윤건[1] · 이지현[2] · 조소희[3]
김문영[3] · 이숭덕[2,3] · 하은호[4]
안재준[4]

[1]연세대학교 응용통계학과
[2]서울대학교 의과대학 법의학교실
[3]서울대학교 의학연구원
법의학연구소
[4]연세대학교 정보통계학과

Received: May 1, 2017
Revised: May 8, 2017
Accepted: May 22, 2017

Correspondence to
Jae Joon Ahn
Department of Information and Statistics, Yonsei University, 1 Yeonsedae-gil, Heungeop-myeon, Wonju 26493, Korea
Tel: +82-33-760-2766
Fax: +82-33-760-2211
E-mail: ahn2615@yonsei.ac.kr

Asian Ethnic Group Classification Model Using Data Mining

Yoon Geon Kim[1], Ji Hyun Lee[2], Sohee Cho[3], Moon Young Kim[3], Soong Deok Lee[2,3], Eun Ho Ha[4], Jae Joon Ahn[4]

[1]Department of Applied Statistics, Yonsei University, Seoul, Korea, [2]Department of Forensic Medicine, Seoul National University College of Medicine, Seoul, Korea, [3]Institute of Forensic Science, Seoul National University College of Medicine, Seoul, Korea, [4]Department of Information and Statistics, Yonsei University, Wonju, Korea

In addition to identifying genetic differences between target populations, it is also important to determine the impact of genetic differences with regard to the respective target populations. In recent years, there has been an increasing number of cases where this approach is needed, and thus various statistical methods must be considered. In this study, genetic data from populations of Southeast and Southwest Asia were collected, and several statistical approaches were evaluated on the Y-chromosome short tandem repeat data. In order to develop a more accurate and practical classification model, we applied gradient boosting and ensemble techniques. To infer between the Southeast and Southwest Asian populations, the overall performance of the classification models was better than that of the decision trees and regression models used in the past. In conclusion, this study suggests that additional statistical approaches, such as data mining techniques, could provide more useful interpretations for forensic analyses. These trials are expected to be the basis for further studies extending from target regions to the entire continent of Asia as well as the use of additional genes such as mitochondrial genes.

Key Words: Y-chromosomal short tandem repeats; Statistical models; Decision trees; Data mining; Ensemble model

서 론

법과학 영역에서 법의유전학이 차지하는 의미는 적지 않다. 법의유전학 영역에서 도입된 소위 '유전자지문'의 개념은 증거물을 통한 감정 결과의 해석에 있어 배제적인 접근에서 결과를 수치적 및 긍정적으로 해석할 수 있는 시작점이 되었다는 점에서 중요하다[1]. 유전자 검사 결과의 수치적 해석은 통계에 기초한다. 통계는 이전부터 친자감별 분야에서 많이 활용되어 왔는데 개인식별 영역에서 variable number tandem repeat, short tandem repeat 유전자들의 활용이 보편화되면서 더욱 중요해졌다. 그리고 다양한 연구 결과의 발표, 폭발적인 정보의 생산 등에 따라 이를 적절하게 해석하여 의미를 부여하고, 실제 사례들에 적용할 수 있기 위해 통계의 중요성은 더욱 강조될 것으로 기대된다.

법의유전학적 접근에서 유전자 검사는 개인식별에 중점을 두고 진행되어 왔다. 최근에는 개인식별 차원을 넘어 좀 더 의미 있는 수사정보를 획득하기 위한 노력을 기울이는 방향으로 연구가 진행되고 있다. 이에는 외형을 예측하려는 노력,

pISSN 2383-5702
eISSN 2383-5710

민족 구분 혹은 출신지 확인 등이 포함되어 있다. 후자의 경우 목적에 따라 대륙 간 구분 혹은 특정 지역의 구분 등으로 범위를 설정할 수 있다. 이러한 연구들은 주로는 민족 간 이동이 큰 유럽을 중심으로 활발하게 진행되어 왔는데, 최근 우리나라를 포함하여 아시아 지역 간에 사회 경제적인 수준이 높아지고 교류가 활발해지면서 아시아 내에서도 민족을 구분할 필요성이 증가하게 되었다.

유전적으로 민족 간 차이가 있음을 뒷받침하는 자료들은 매우 많다[2-4]. 법과학적으로 민족 간 유전적 차이를 활용하면 특정 유전자형의 사람이 어디에서 왔는지를 비교적 정확하게 구분해 낼 수 있어 유용하다. 유전적 차이는 흔히 비교 대상이 되는 민족들을 대상으로 기본 자료를 획득하고, 이러한 차이를 비교 분석하여 목적에 따라 적절하게 가공하여야 한다. 흔히 지도상에 특정 유전자형(혹은 일배체형)이 어떻게 분포하는지를 빈도 혹은 농도로 환산하여 표시하는 방법이 있다[5]. 하지만 실무에서는 단순히 민족 간 차이가 있다는 사실 이외에도 차이를 어떻게 활용할 수 있는지 좀 더 구체적인 접근을 필요로 하게 된다. 예를 들어, 특정 유전자형의 사람이 A라는 지역에서 왔을 가능성이 B라는 지역에서 왔을 가능성과 비교하여 어떠한지 등과 같이 결과 혹은 예측을 수치화하여 유전자 검사 결과를 사용하는 실무자들에게 좀 더 편하고 유용하게 결과를 제공하려는 시도가 필요하다. 이러한 시도들은 다양하게 진행되고 있는데, 아직 일상적이라고만은 할 수 없고, 또 어떠한 방법이 좋은지 혹은 방법에 따른 차이는 어떠한지 등과 관련한 정보들은 그리 충분하다고 할 수 없다. 이번 연구에서는 (1) 민족 간 차이가 비교적 큰 Y-chromosome short tandem repeat (Y-STR)를 중심으로, (2) 동남아시아와 서남아시아인의 데이터를 이용하여 특정인을 두 개 지역 중 하나로 예측할 때 (3) 다양한 통계적 접근방법에 따라 어떠한 차이가 나오는지를 확인하고, (4) 이러한 접근에서 중요한 여러 통계적 방법에 대한 기본적인 지식을 익힘으로써, 법의유전학 분야에서의 통계의 활용을 위한 이해를 높이고, 그 중요성을 다시 한 번 확인해 보고자 한다.

본 론

1. 분류분석 (classification analysis)

본 연구에서는 민족구분을 위한 분류분석을 위해 데이터마이닝 분류 알고리즘을 사용하였다. 분류분석이란 소속집단을 알고 있는 데이터를 가지고 집단을 분류하는 모형을 학습시킨 이후 소속집단을 모르는 데이터를 어느 특정 소속집단으로 분류하는 방법이다. Fig. 1은 분류분석 프로세스를 도식화한 것이다.

분류분석은 어떤 종류의 분류함수를 사용하는지에 따라 다양한 종류의 분류모형들이 존재하며, 대표적으로 많이 사용되는 분류모형으로는 의사결정나무(decision tree) 모형, 로지스틱회귀(logistic regression) 모형 그리고 신경망(neural network) 모형 등이 존재한다.

분류모형은 입력변수 x_{ij}가 주어졌을 때 추정값 및 예측값 y_i를 도출하는 방식으로 다음의 Eq. (1)과 같이 입력변수 들의 선형결합으로 나타낼 수 있다.

$$\hat{y}_i = w_1 x_{i1} + w_2 x_{i2} + \cdots + w_k x_{ik} = \sum_{j=1}^{k} w_j x_{ij} \quad (1)$$

위의 Eq. (1)은 x_{ij}들의 선형식으로 표현이 되어 있지만 특정 함수들의 식으로도 표현 가능하다.

(1) 의사결정나무 모형

의사결정나무 모형은 데이터마이닝의 분류모형 중 가장 대표적인 분류모형이다. 의사결정나무분석은 의사결정규칙(rule)을 나무구조에 의한 추론규칙으로 표현하여 분류 및 예측을 하는 분석 방법으로 목표(target)변수가 이산형 목표변수인 경우를 의사결정나무라고 하며, 연속형 목표변수인 경우를 회귀나무라고 한다. 의사결정나무는 크게 노드(node)와 가지(branch)로 이루어져 있으며, 분리기준(split criterion), 정지규칙(stopping rule)을 지정하여 나무를 만들게 되며, 분류오류(classification error)의 위험성과 추론규칙(induction rule)의 적절성을 판단하여 불필요한 가지를 제거해 나가며 형성하게 된다. 나무구조의 의사결정규칙을 형성할 때 최적

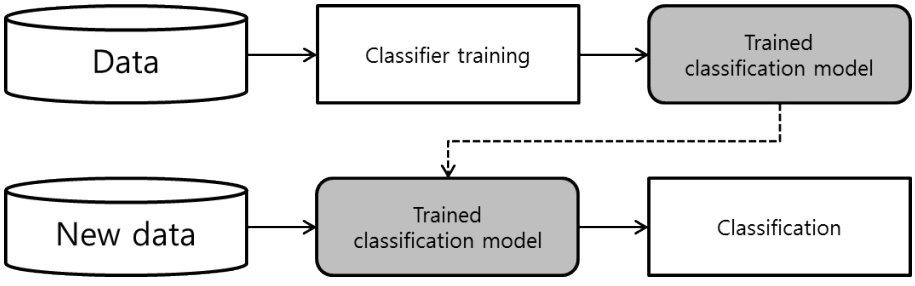

Fig. 1. Classification analysis process.

의 분리기준 결정을 위해 적용되는 통계량으로 카이제곱(χ^2) 통계량, 지니(Gini) 계수, 그리고 엔트로피(entrophy) 계수가 있다. 카이제곱 통계량은 의사결정나무에서 분리규칙을 형성할 때 각 입력변수에 대해 범주집단의 분포가 독립적인가를 검정하여, 각 유의확률(P-value)이 가장 작은 값을 보이는 입력변수를 선택하여 최적의 분류값을 찾아 가지와 노드를 형성해나가는 방식을 보인다. 지니 계수는 불순도 혹은 불균형의 정도를 측정하는 지수로서 지니 계수의 값이 0에 가까울수록 균등하며 1에 가까울수록 불균등함을 의미한다. 따라서 지니 계수를 이용하여 분리규칙을 형성할 때는 지니 계수를 가장 감소시켜주는 입력변수들을 선택하여 최적분리를 진행하게 된다. 엔트로피 계수는 지니 계수와 유사한 계수로 불순도와 관련된 지수이며, 엔트로피 계수를 최소로 만드는 입력변수를 찾아내어 최적의 분류값을 도출해나가며 분리규칙을 형성해나간다[6]. Fig. 2는 고객의 구매 및 비구매 분류에 대한 의사결정나무의 의사결정규칙을 그림으로 표현한 예이다.

(2) 앙상블 (ensemble) 모형

앙상블 모형이란 동일한 종류의 분류모형들 혹은 서로 상이한 종류의 분류모형들의 예측 결과들을 결합하여 최종적인 의사결정을 내리는 모형을 말한다. 이론적으로는 앙상블 모형을 사용한 분류모형이 단일기반처리에 기반한 분류모형보다 예측 오분류율이 낮은 것으로 알려져 있다[7-9]. 단일기반처리란 한 가지 분류기를 이용하여 분류모형을 만드는 것을 의미한다. 예를 들어 동일한 데이터에 대해 상호 독립적인 k개의 분류기를 사용하여 만든 k개의 분류모형들이 모두 P의 오분류율을 가지고 있다고 하자. 각 분류기가 서로 상호 독립적인 모형이기에, 각 분류기의 절반 이상이 오분류를 할 경우에 대한 앙상블 모형의 오분류율은 다음과 같다.

$$e_{Ensemble} = \sum_{i=[k/2]}^{k} \binom{k}{i} P^i (1-P)^{k-i} \quad (2)$$

Eq. (2)의 k와 P에 특정한 값을 넣어서 계산을 하게 되면 오분류율 $e_{Ensemble}$은 P보다 낮게 나오게 되는 것을 알 수 있다.

1) 배깅 (bagging) 기법

배깅은 bootstrap aggregating의 줄임말이다. 배깅은 앙상블 모형을 구축할 때 사용할 수 있는 기법으로, 훈련용 데이터 셋으로부터 붓스트랩 자료를 k번 반복추출하여 k개의 붓스트랩 자료(표본)를 만들고, 각 표본에 대해 분류모형을 만들어 결과를 종합하여 분류결과를 도출하게 된다[10]. Fig. 3은 배깅 절차를 도식화한 것이다.

2) 부스팅 (boosting) 기법

부스팅은 데이터를 학습시키는 과정에서 예측력이 약한 예측모형들을 결합하여 최종적인 분류모형의 예측력을 향상시키는 기법으로 배깅과 유사한 기법이다. 그런데 배깅은 k개의 붓스트랩 자료에 분류기들을 학습시킬 시에 각 분류기들이 상호 영향을 주지 않는다는 점이 존재하는 반면, 부스팅의 경우 k개의 붓스트랩 자료를 순차적으로 분류기에 학습시키며 이전 분류기의 학습 결과가 이후의 붓스트랩 자료를 만들 때 영향을 준다. 부스팅에서 표본을 추출할 때 이전 분류기의 분류결과에서 잘못 분류한 데이터와 학습에 이용되지 않은 데이터에 가중치를 주어 예측력이 약한 데이터와 학습시키지 않은 데이터를 더 잘 훈련시켜 예측력이 강한 분류모형을 구축하게 된다[11,12]. Fig. 4은 부스팅 절차를 도식화한 것이다.

3) 그래디언트 부스팅 (gradient boosting)

그래디언트 부스팅은 일반화 부스팅의 종류로서 기울기 강하

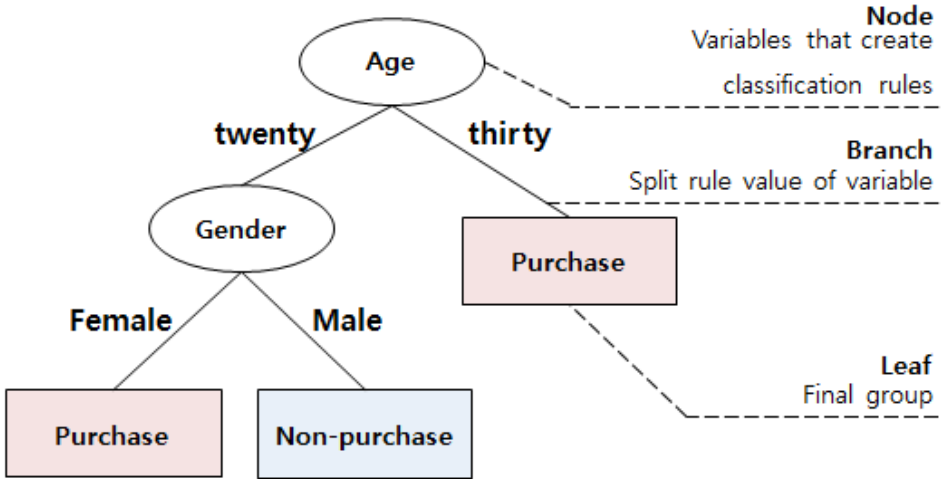

Fig. 2. Examples of decision rules.

https://doi.org/10.7580/kjlm.2017.41.2.32

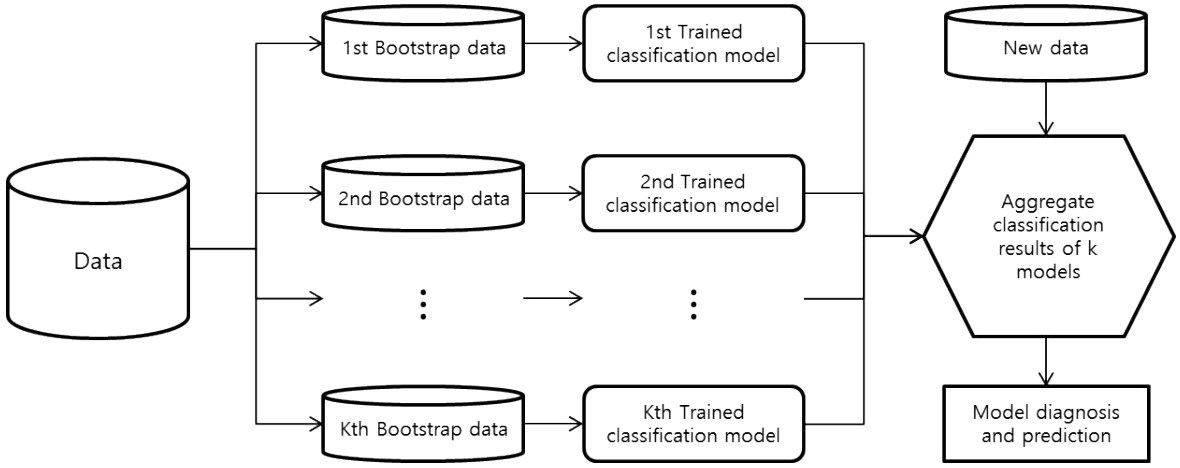

Fig. 3. Bagging procedure.

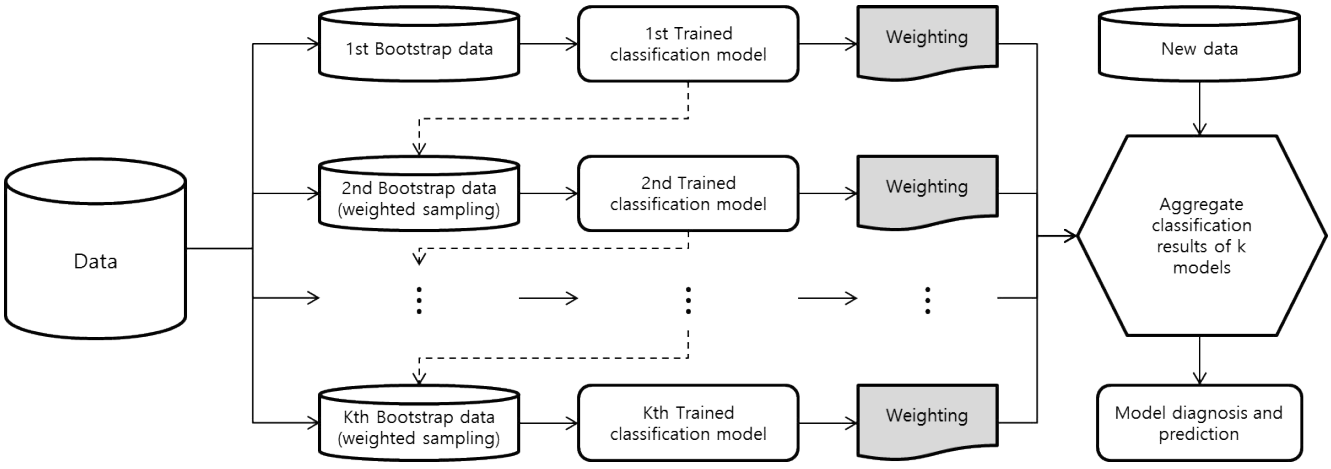

Fig. 4. Boosting procedure.

(gradient descent) 알고리즘을 이용하여 분류모형의 기대손실을 최소화시키는 방향을 찾아나가게 된다[13]. 기울기 강하 알고리즘의 아이디어는 뉴턴-랩슨(Newton-Raphson)법과 연관이 깊다.

2. 샘플링 (sampling)

분류모형을 학습시킬 때 이항형 목표(binary target)변수의 비율이 2:8 혹은 1:9와 같이 한 범주 값에 비율이 상당히 치우쳐져 있는 경우를 계급불균형(imbalanced) 자료라 부른다. 계급불균형 자료의 원 비율을 가지고 분류모형을 구축할 시 분류모형이 자료를 모든 자료를 비율이 높은 범주로 분류하게 되어도 분류 정확도가 상당히 높게 나오게 되는데 이는 신뢰하기 어려운 결과로 볼 수 있다. 따라서 분류모형 결과의 신뢰성 문제를 해결하기 위해 샘플링 기법을 통해 계급불균형 자료의 계급비를 5:5와 같은 균형비로 맞추어 주어 분류모형을 학습시키는 과정이 필요하다[14,15]. 대표적으로 많이 사용되는 샘플링 기법으로는 언더샘플링(under sampling) 기법과 오버샘플링(over sampling) 기법이 존재하며, 본 연구에서는 언더샘플링 기법을 사용하였다. 언더샘플링 기법은 Fig. 5와 같이 다수의 집단을 비복원 추출 방법을 이용하여 다수의 비율을 낮추어주어 소수의 집단의 비율에 맞추는 방법이다. 실제 데이터만을 가지고 분석을 진행한다는 장점이 존재하지만, 원 자료를 손실한다는 단점이 있다.

3. 실험 방법

(1) 데이터 수집

본 연구에는 서울대 법의학교실이 보유한 데이터(베트남인, 인도인, 네팔인)와 논문을 통해 공개된 Y-STR 데이터를 사용하였다[16]. PowerPlex Y23 System (Promega Corporation, Madison, WI, USA)을 이용하여 유전자 검

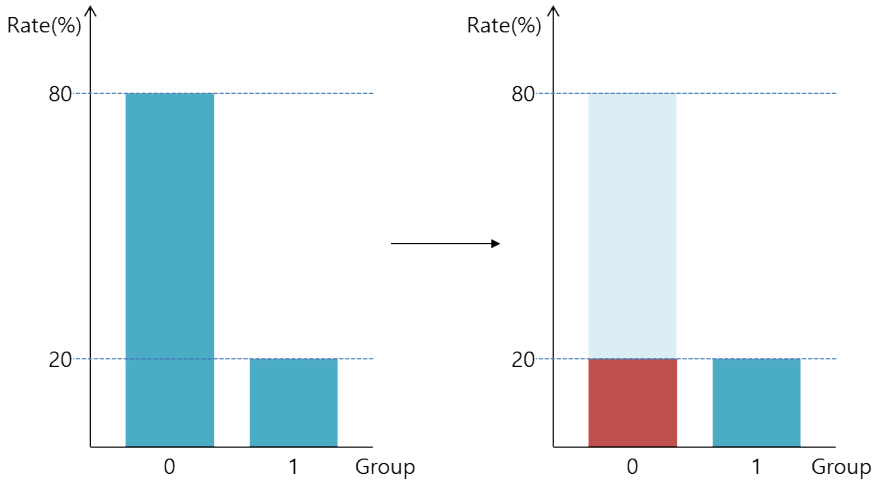

Fig. 5. Under sampling.

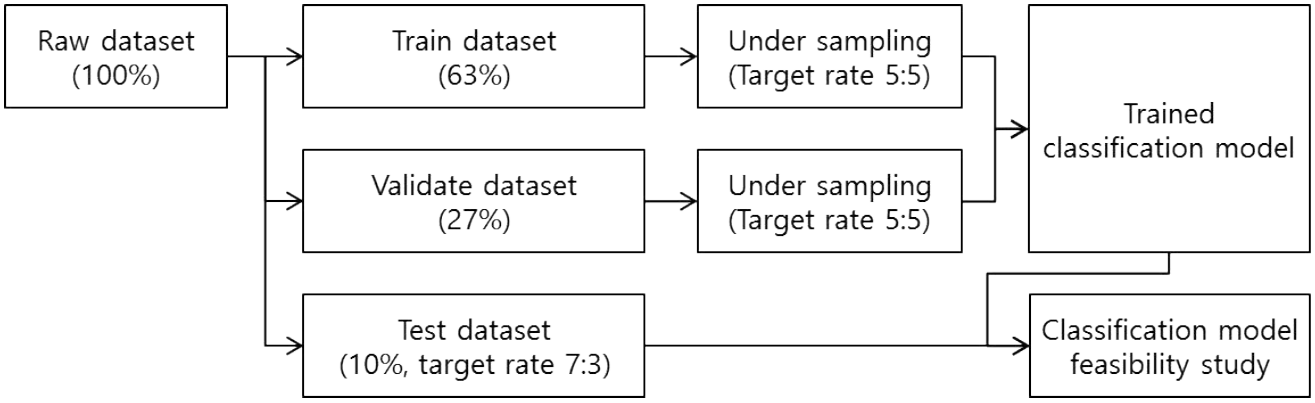

Fig. 6. Progress of ethnicity classification model analysis.

사를 진행한 데이터로 총 1,345명의 데이터를 수집하였다 (Table 1). 본 데이터 수집에 대한 사항은 서울대학교 의과대학 및 서울대병원 Institutional Review Board (IRB) 심의 과정을 거쳤다.

(2) 데이터 구성

본 자료의 변수는 Table 2와 같이 국가 정보와 21가지의 유전자변수 그리고 동남아시아인과 서남아시아인을 구분하는 목표변수를 포함하여 총 23가지의 변수들로 구성되어 있다.

(3) 데이터마이닝 분석진행과정

Y-STR 자료를 이용하여 동남아시아인과 서남아시아인을 분류하는 모형을 만들기 위해 본 연구에서 분석을 진행한 과정은 Fig. 6과 같다.

Table 1. Details of populations analyzed

Population	Sample size	Data source
Vietnam	46	Seoul National University
Nepal	69	
India	23	
Vietnam	45	Purps et al. [16]
Philippines	798	
Singapore	104	
India	298	
Total	1,383	

1) 데이터 분할 및 샘플링

데이터 분할 과정은 모형을 구축하고, 구축한 모형을 평가하기 위해 데이터 셋을 훈련용(train), 평가용(validation), 검증용(test)으로 나누는 것을 말한다. 본 분석에서는 Fig. 7과 같

https://doi.org/10.7580/kjlm.2017.41.2.32

이 원 데이터 셋을 훈련용, 평가용, 검증용으로 나누는 작업을 하였으며, 데이터 분할 진행 과정은 다음과 같다.

1) Y-STR데이터 셋을 훈련용 및 평가용(90%)과 검증용(10%)으로 분할
2) 90%의 훈련용 및 평가용 데이터셋을 훈련용(70%), 평가용(30%)으로 분할

본 연구에서는 동남아시아인과 서남아시아인의 비율은 약 72:28로 계급불균형 자료이다. 따라서 본 연구에서 사용하는 유전자 데이터의 특성상 온전한 실제 데이터만 사용하는 것이 적절하다고 판단하여 언더샘플링 기법을 통해 훈련용 데이터 셋과 평가용 데이터 셋의 계급비를 5:5로 맞추어 주어 분석을 진행하였으며, 검증용 데이터 셋의 계급비는 원 비율을 사용하여 모형을 검증하였다.

2) 분류분석모형의 적용

데이터 분할 및 샘플링 기법을 적용한 이후 훈련용 및 평가용 데이터 셋에 대하여 데이터마이닝의 분류모형을 학습시키는 과정을 거쳤다. 분류분석의 기법으로는 의사결정나무 모형,

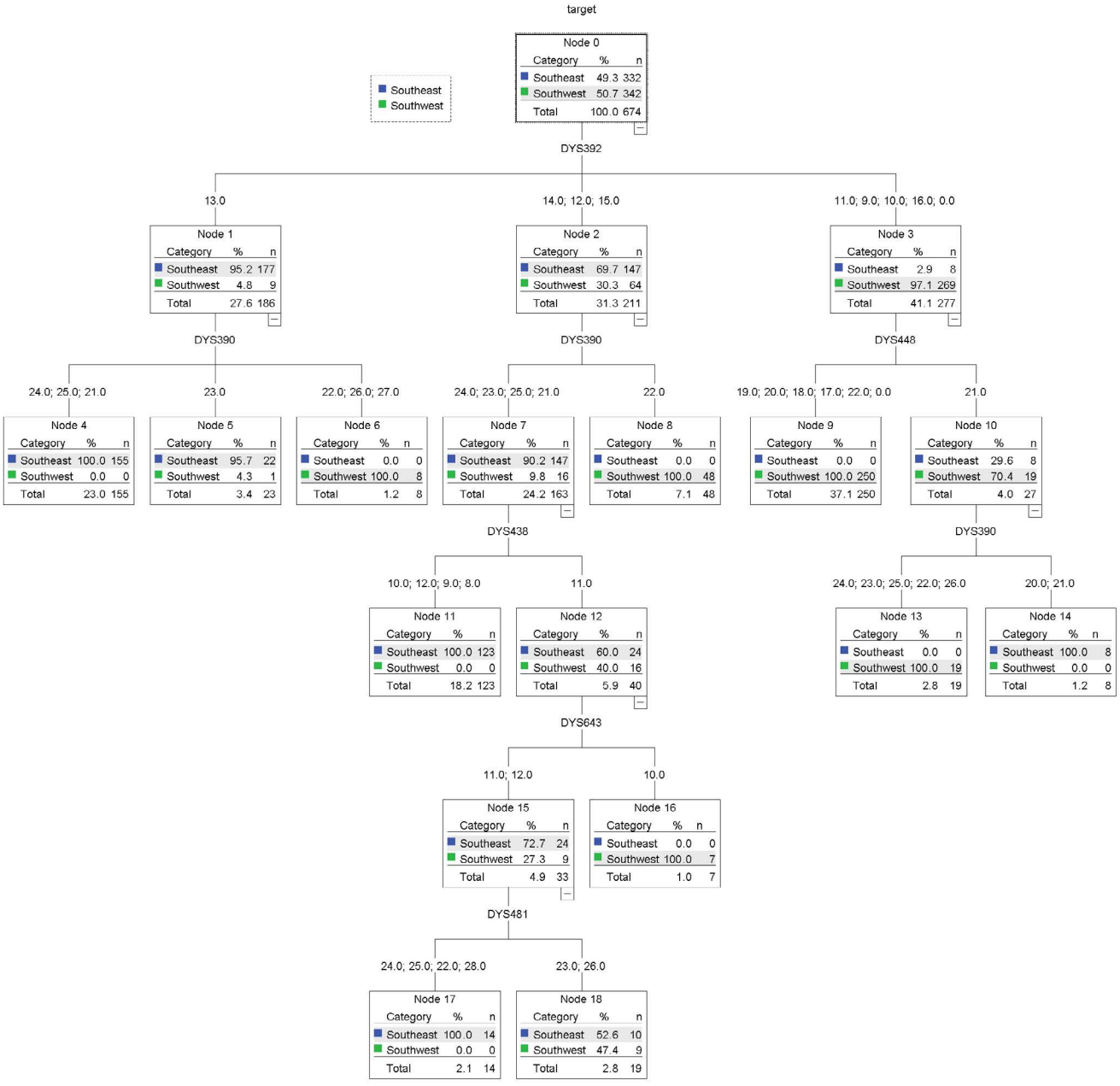

Fig. 7. Gradient boosting and decision tree (chi-square) ensemble model separation rule tree.

배깅 및 부스팅과 같은 리샘플링 기법, 앙상블 기법 그리고 그래디언트 부스팅 모형을 사용하여 총 22가지의 분류모형들을 구축하였다. 그리고 학습된 22가지의 모형들을 검증용 데이터 셋을 이용하여 모형의 타당성 및 모형성능을 검증하는 과정을 거쳤다.

본 연구의 22가지의 분류모형들에서 단일기반처리인 의사결정나무 모형 3가지를 제외하고는 모두 앙상블 모형이다. 앙상블 모형은 블랙박스(black box) 모형으로 학습된 모형의 분류결과를 해석하기 어렵다는 단점이 존재한다. 본 연구에서는 분류모형들의 결과가 어떻게 도출되는지 설명하기 위해 단일기반처리로 학습시킨 의사결정나무 모형과 앙상블 모형의 결과를 의사결정나무 모형에 다시 학습시켜 분류결과를 해석하는 과정을 거쳤다.

4. 결과

(1) 데이터 분할 및 언더샘플링

본 연구에 사용된 데이터는 1,345개의 개체 수를 가지고 있으며, 목표변수인 동남아시아인과 서남아시아인의 비율은 약 72대 28인 계급불균형 자료이다. Table 3은 데이터 분할과 언더샘플링 결과에 대해 보여주고 있다.

Table 2. Composition of data

No.	Variable	Definition
1	Sample Info	National information
2	DYS576	Gene
3	DYS389I	
4	DYS448	
5	DYS389II	
6	DYS19	
7	DYS391	
8	DYS481	
9	DYS549	
10	DYS533	
11	DYS438	
12	DYS437	
13	DYS570	
14	DYS635	
15	DYS390	
16	DYS439	
17	DYS392	
18	DYS643	
19	DYS393	
20	DYS458	
21	DYS456	
22	YGATAH4	
23	TARGET	0: Southeast Asian 1: Southwest Asian

(2) 분류분석의 적용 결과

본 연구에서는 동남아시아인과 서남아시아인을 분류하는 예측모형을 최적화시키기 위하여 의사결정나무 모형과 리샘플링 기법인 배깅 및 부스팅 그리고 앙상블 기법을 사용하여 총 22가지의 분류모형을 구축하였다. Table 4는 총 22가지의 분류모형의 오분류율과 receiver operation characteristic (ROC) 인덱스에 대한 결과이다. ROC는 여러 절단값에서 민감도와 특이도의 관계를 2차원의 좌표평면 위에 곡선의 형태를 보여주어 분류기의 성능을 시각화시킨 것으로, 곡선의 넓이 부분을 계산하여 분류기의 성능을 평가한다. 이러한 곡선의 넓이를 ROC 인덱스 혹은 area under the curve라고 하며, ROC 인덱스 통계량 값이 클수록 분류기의 성능이 우수하다고 할 수 있다. 결과를 보면 22가지 분류모형 대부분의 분류 성능이 우수하게 나왔으며, 검증용 데이터 셋에 대한 평균 오분류율은 약 6%이었다. 또한 22가지의 모형 중 단일기반처리에 의한 분류모형보다 앙상블 모형들의 오분류율 및 ROC 인덱스가 우수하게 나온 것을 확인할 수 있다. 특히 그래디언트 부스팅과 의사결정나무의 앙상블 모형의 검증용 오분류율이 약 4%대를 보이고, ROC 인덱스 또한 약 0.99의 값을 보여 동남아시아인과 서남아시아인을 분류에 있어서 뛰어난 예측성능을 보였다. 분류모형의 결과가 어떤 규칙을 기반으로 도출되었는지 살펴보기 위해 Table 4에서 가장 분류성능이 우수하게 나온 1번 모형인 그래디언트 부스팅과 의사결정나무(chi-square) 모형의 앙상블 모형의 분류결과를 의사결정나무에 학습시켰다.

1) 그래디언트 부스팅 및 의사결정나무의 앙상블 모형 분류결과

Table 5는 그래디언트 부스팅과 의사결정나무(chi-square)형의 앙상블 모형의 분류결과를 의사결정나무에 학

Table 3. The results of data splitting and under sampling

Category	Dataset	Count	Target rate
Raw data	Y_STR_Raw	1,345	72.0:28.0
Data partition	Train dataset	846	72.0:28.0
	Validate dataset	364	72.0:28.0
	Test dataset	135	72.0:28.0
Under sampling	Train dataset (under sampling)	470	50:50:00
	Validate dataset (under sampling)	204	50:50:00

https://doi.org/10.7580/kjlm.2017.41.2.32

습시킨 결과이다. Table 5를 보면 입력변수에 사용된 유전자 변수 21개 중 5개의 변수들이 분류모형을 구축하는 데 중요하게 사용된 것을 볼 수 있다. 특히 DYS392와 DYS390은 다른 변수들보다 민족을 구분하는 데 중요한 변수들로 사용되는 것을 볼 수 있다. Fig. 7은 그래디언트 부스팅 및 의사결정나무의 앙상블 모형의 분류결과를 의사결정나무에 다시 학습시켜 도출된 분류규칙나무이다.

고 찰

최근 들어 유전자 검사를 통해 단순한 개인식별의 차원을 넘어 의미 있는 수사정보를 획득하려는 움직임이 활발하고, 그 가운데에는 유전자가 어느 지역의 사람으로부터 유래되었는지 여부를 확인하는 것도 포함되어 있다. 특히 아시아 지역에서도 역시 세계화에 따른 활발한 교류 덕분에 이러한 필요성은 꾸준히 증가하고 있다. 이런 면에서 본 연구는 매우 시의

Table 5. Ensemble model variable importance

Variable	Count of split rules	Variable importance	
		Train	Validate
DYS392	1	1	1
DYS390	2	0.68	0.649
DYS448	1	0.256	0.193
DYS643	1	0.219	0.13
DYS438	1	0.193	0.279

Table 4. Result of classification model

No.	Model	Resampling	Misclassification rate			ROC index		
			Train	Validate	Test	Train	Validate	Test
1	GB and DT (Chi-square) Ensemble	Bagging	0.038	0.068	0.044	0.996	0.975	0.992
2	GB and DT (Chi-square) Ensemble	Boosting	0.044	0.063	0.037	0.995	0.973	0.99
3	DT (Entropy) and DT (Entropy) Ensemble	Bagging and boosting	0.046	0.073	0.037	0.995	0.978	0.992
4	GB and DT (Gini) Ensemble	Boosting	0.046	0.078	0.037	0.995	0.968	0.992
5	DT (Gini) and DT (Gini) Ensemble	Bagging and boosting	0.055	0.092	0.037	0.993	0.969	0.994
6	DT (Chi-square) and DT (Chi-square) Ensemble	Bagging and boosting	0.057	0.083	0.037	0.993	0.977	0.992
7	GB and DT (Entropy) Ensemble	-	0.063	0.087	0.044	0.98	0.966	0.985
8	GB and DT (Gini) Ensemble	-	0.063	0.087	0.044	0.98	0.966	0.985
9	GB	-	0.065	0.063	0.044	0.981	0.966	0.984
10	GB and DT (Chi-square) Ensemble	-	0.067	0.087	0.052	0.98	0.966	0.984
11	GB and DT (Entropy) Ensemble	Bagging	0.069	0.083	0.037	0.981	0.972	0.987
12	DT (Gini)	-	0.069	0.083	0.052	0.95	0.955	0.969
13	DT (Entropy)	-	0.069	0.083	0.052	0.95	0.955	0.969
14	GB and DT (Gini) Ensemble	Bagging	0.071	0.078	0.037	0.98	0.971	0.988
15	GB and DT (Chi-square) Ensemble	Bagging	0.071	0.087	0.037	0.98	0.971	0.988
16	DT (Chi-square)	-	0.076	0.083	0.059	0.948	0.954	0.967
17	DT (Chi-square)	Bagging	0.08	0.073	0.067	0.963	0.97	0.989
18	DT (Gini)	Bagging	0.08	0.073	0.067	0.963	0.97	0.989
19	DT (Entropy)	Bagging	0.084	0.083	0.059	0.973	0.966	0.983
20	DT (Gini)	Boosting	0.137	0.248	0.163	1	0.962	0.993
21	DT (Chi-square)	Boosting	0.149	0.15	0.126	1	0.973	0.993
22	DT (Entropy)	Boosting	0.179	0.238	0.185	1	0.977	0.991

ROC, receiver operation characteristic; GB, gradient boosting; DT (Chi-square), decision tree model using chi-square statistics; DT (Entropy), decision tree model using chi-square (entropy) statistics; DT (Gini), decision tree model using chi-square (Gini) statistics.

적절하다고 하겠다.

지역간 차이는 본 연구에서와 같은 Y 염색체 연구뿐만 아니라 모계로 유전되는 미토콘드리아 유전자 연구, 그리고 일상적인 상염색체 유전자 연구 등 다양한 방법으로 진행할 수 있다. 본 연구 결과에서도 보는 바와 같이 지역간 차이를 넘어 이에 기초하여 예측 가능한 모델을 만들기 위해서는 다양한 통계적인 접근이 필요할 수 있고, 또 이에 필요한 여러 자료들은 쉽게 얻을 수 있어야만 한다. 이런 측면에서 Y 염색체 검사 결과에 기반한 연구는 민족간 다양성이 상대적으로 크고, 비교적 쉽게 자료를 얻을 수 있으며, 이미 공개된 자료도 얻을 수 있다는 점 등에서, 예측 모델 작성 연구의 시작점으로 적정하다고 하겠다. 그리고 본 연구결과를 통해 얻은 다양한 접근 방법에 대한 결과는 향후 좀 더 많은 지역을 대상으로 한 연구, 그리고 다른 유전자를 활용한 연구에서도 매우 중요한 경험을 제공하여 줄 것으로 기대한다.

본 연구에서 결과는 좀 더 구체적으로 수치화하여 제시할 수 있었고, 나아가 그 과정을 시각화 할 수도 있어 결과 설명에 있어서도 매우 의미 있다고 하겠다. 한편 어느 방법을 사용하느냐 여부에 따라 결과는 다소 차이가 있었다. 예를 들면 일부 결과에서는 10% 넘는 오분류율을 보이기도 하였다. 본 연구에서와 같은 접근이 일반적이지 않음을 고려할 때 위와 같은 수치가 어느 정도 의미가 있는지, Y 염색체 검사 결과를 활용한 연구의 원래 제한점은 아닌지, 혹은 좀 더 넓은 민족을 대상으로 하였을 때 그러한 수치는 어떻게 변할 것인지 등에 대해 구체적인 의견을 주기는 어려운 상황이다. 좀 더 많은 대상으로 한 연구가 필요함을 시사한다.

본 연구 이전에 진행되었던 의사결정나무 모형이나 회귀모형과 비교하여 분류모형들의 전반적인 성능이 동남아시아 및 서남아시아 사람들을 분류하는 데 있어 우수한 성능을 보였다. 그리고 어느 방법을 사용하는지 여부에 따라 오분율 등의 결과에 있어서는 차이가 있었다. 결국 데이터마이닝 기법 등의 다양한 접근은 매우 중요하고, 나아가 대상 지역이 넓어지고 추가적인 데이터베이스가 구축될수록 가장 효율적인 방법은 변할 가능성도 배제할 수 없겠다. 결국 이를 위해서는 통계 전문가와 법의유전학자들 사이에 꾸준한 협력과 지속적인 자료의 보완이 중요할 것임을 시사한다.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This research was supported by the Bio and Medical Technology Development Program of the National Research Foundation (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2014M3A9E1069989).

References

1. Butler JM. Advanced topics in forensic DNA typing: methodology. San Diego, CA: Academic Press; 2011.
2. Enoch MA, Shen PH, Xu K, et al. Using ancestry-informative markers to define populations and detect population stratification. J Psychopharmacol 2006;20;(4 Suppl):19-26.
3. Li JZ, Absher DM, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science 2008;319:1100-4.
4. Rosenberg NA, Pritchard JK, Weber JL, et al. Genetic structure of human populations. Science 2002;298:2381-5.
5. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155:945-59.
6. Quinlan JR. Induction of decision trees. Mach Learn 1986;1:81-106.
7. Opitz D, Maclin R. Popular ensemble methods: an empirical study. J Artif Intell Res 1999;11:169-98.
8. Rokach L. Ensemble-based classifiers. Artif Intell Rev 2010;33:1-39.
9. Quinlan JR. Bagging, boosting, and C4.5. In: AAAI/IAAI '96 Proceedings of the Thirteenth National Conference on Artificial Intelligence; 1996 Aug 4-8; Portland, OR, USA. Vol. 1. Palo Alto, CA: AAAI Press; 1996. p. 725-30.
10. Breiman L. Bagging predictors. Mach Learn 1996;24:123-40.
11. Schapire RE. The strength of weak learnability. Mach Learn 1990;5:197-227.
12. Freund Y, Schapire RE. A short introduction to boosting. J Jpn Soc Artif Intell 1999;14:771-80.
13. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002;38:367-78.
14. Wang R, Lee N, Wei Y. A case study: improve classification of rare events with SAS Enterprise Miner. In: Proceedings of the SAS Global Forum 2015 Conference. Cary, NC: SAS Institute Inc.; 2015.
15. Rahman MM, Davis DN. Addressing the class imbalance problem in medical datasets. Int J Mach Learn Comput 2013;3:224-8.
16. Purps J, Siegert S, Willuweit S, et al. A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. Forensic Sci Int Genet 2014;12:12-23.