The Korean Journal of Applied Statistics (2018) DOI: http://dx.doi.org/10.5351/KJAS.2018.31.3.417
31(3), 417–426

Comparison of the performance of classification algorithms using cytotoxicity data

Yeochang Yoon[a] · Eui Bae Jeung[b] · Na Rae Jo[c] · Su In Ju[c] · Sung Duck Lee[c,1]

[a]Department of Information Security, Woosuk University;
[b]Department of Veterinary Medicine, Chungbuk National University;
[c]Department of Statistics, Chungbuk National University

(Received May 28, 2018; Revised June 7, 2018; Accepted June 8, 2018)

Abstract

An alternative developmental toxicity test using mouse embryonic stem cell derived embryoid bodies has been developed. This alternative method is not to administer chemicals to animals, but to treat chemicals with cells. This study suggests the use of Discriminant Analysis, Support Vector Machine, Artificial Neural Network and k-Nearest Neighbor. Algorithm performance was compared with accuracy and a weighted Cohen's kappa coefficient. In application, various classification techniques were applied to cytotoxicity data to classify drug toxicity and compare the results.

Keywords: discriminant analysis, support vector machine, artificial neural network, k-nearest neighbors classification, weighted Cohen's kappa coefficient, cytotoxicity

1. 서론

기존의 약물독성시험법은 동물실험을 통해 이루어져 왔는데, 이는 경제적 측면과 시간적 측면, 생명윤리적 측면에서 많은 문제점을 가지고 있다. 동물대체시험법 중 하나로 사용되는 쥐의 배아줄기세포 시험법(mouse embryonic stem cell test; mEST)은 유럽대체실험검증센터(European Centre for the Validation of Alternative Methods; ECVAM)에서 제안한 방법으로, 미분화된 쥐의 배아줄기세포(mouse embryonic stem cells; mESC) 및 분화된 쥐의 섬유아세포에 약물을 주입한 후 생존력이 50%가 되는 시점의 약물의 농도와 mESC의 심근세포로의 분화가 50%가 되는 시점의 약물의 농도를 측정하여 약물의 독성을 평가한다.

기존의 연구로 Genschow 등 (2000)은 선형판별분석을 이용해 mEST방법과 기타 다른 방법을 비교하였고, Seiler 등 (2004)은 개선된 mEST방법을 통해 선형판별분석을 이용해 약물의 독성을 분류한 결과, 78%의 분류 정확도를 얻었다.

미분화된 쥐의 배아줄기세포에서 심근분화를 유도하기 위해서는 배상체(embryoid body) 형성을 반드시 필요로 하는데 Kang 등 (2017)은 이 과정에서 발생독성물질의 농도 증가에 따라 배상체의 단면적

This research was supported by a grant (17182MFDS487) from Ministry of Food and Drug Safety in 2017.
[1]Corresponding author: Department of Statistics, Chungbuk National University, 1, Chungdae-ro, Seowon-gu, Cheongju, Chungbuk 28644, Korea. E-mail: sdlee@chungbuk.ac.kr

이 감소하는 것을 발견하였다. 배상체 단계에서 배상체 단면적을 통해 약물 독성을 평가할 경우, 심근분화 유도에 소요되는 시간과 노동력을 감소시킬 수 있다는 이점이 있다. 따라서 Kang 등 (2017)은 기존의 mEST방법에서 측정하는 값 중 하나인 심근세포로의 분화가 50%가 되는 시점에서의 약물의 농도를 쥐의 배상체의 단면적이 50%가 되는 시점에서의 약물의 농도로 대체하여 좀 더 개선된 배상체 단면적을 활용한 시험법(mouse embryoid body test; mEBT)을 개발하였으며, 그 결과, 90.5%의 분류 정확도를 얻었다. 최근 머신러닝의 발전으로 다양한 분류 알고리즘을 적용할 수 있게 되면서 본 연구에서는 통계적 기법(판별분석)과 머신러닝 기법(서포트 벡터 머신(support vector machine; SVM), 인공신경망(artificial neural network; ANN), k-인접이웃분류(k-nearest neighbor; k-NN))의 성능을 비교하여 최적의 분류 알고리즘을 제안하고자 한다.

Yu (2010)는 마이크로 어레이 데이터를 활용하여 분류 알고리즘의 성능을 비교하였고 그 결과, 서포트 벡터 머신이 가장 좋은 성능을 나타내었다.

본 연구는 세포독성 자료로 3단계의 발생독성(무독성, 중간 독성, 강한 독성)을 예측하였으며, 분류 예측 방법에 따라 성능을 비교하는 과정을 주 내용으로 한다. 2장에서는 분류 알고리즘의 방법론과 모형 평가 기준에 대해서 설명하였으며, 3장에서는 실증분석으로 실제 세포독성 자료로 여러 가지 분류 기법들을 적용하여 약물의 독성을 분류(예측)하였고 그 결과를 비교하였다. 마지막으로 4장에서는 연구에 대한 결론을 내리고 추후 연구에 대해 언급하였다.

2. 분류분석 방법론

2.1. 이차판별분석

판별분석에서는 모집단에 대한 다변량 정규성, 그룹내 공분산행렬의 동일성, 변수들 간의 낮은 다중공선성 가정이 요구된다. 이러한 가정하에서 새로운 객체에 대하여 사후확률을 최대로 하는 그룹으로 분류한다.

선형판별분석과 달리 이차판별분석(quadratic discriminant analysis; QDA)은 조건부분포 $\pi(x|y = C_k)$에 대해 그룹-특정적(class-specific) 평균벡터 μ_k과 개별 공분산행렬 Σ_k를 가지는 다변량 정규분포 $N(\mu_k, \Sigma_k)$를 가정한다. 이때, 이차판별함수는

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k \\ &= -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + \log\pi_k \end{aligned} \tag{2.1}$$

으로 주어지고, 선형판별분석과 마찬가지로 새로운 객체에 대해 이차판별함수의 값(식 (2.1))이 가장 큰 그룹으로 분류한다.

2.2. 서포트 벡터 머신

Vapnik (1998)에 의해서 정립된 서포트 벡터 머신은 기계학습(machine learning)의 한 방법으로, 훈련용 데이터를 비선형 매핑을 통해 고차원으로 변환한다. 이 새로운 차원에서 하나의 그룹을 다른 것으로부터 분리하는 의사결정 영역인 초평면(hyperplane)을 찾고, 찾아낸 초평면을 이용하여 새로 유입되는 데이터들을 분류하게 된다. 서포트 벡터 머신은 두 개의 그룹이 있는 경우에 개발된 것으로 두 개의 그룹을 $+1$과 -1로 구분하여 분류하게 된다. 초평면 함수 $f(x)$는 p-차원에서 선형판별함수를 제공하며, 원래의 공간을 2개의 그룹을 분리할 수 있는 공간으로 나눈다.

$$f(x) = w^Tx + b = w_1x_1 + w_2x_2 + \cdots + w_px_p + b. \tag{2.2}$$

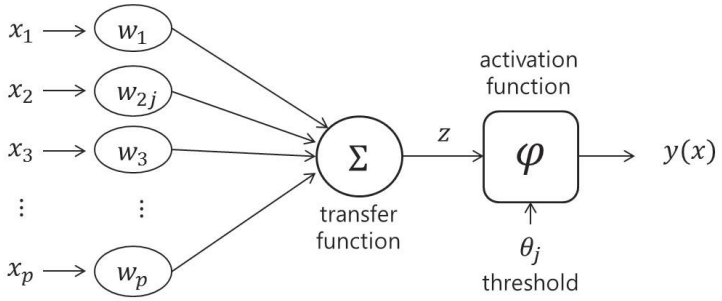

Figure 2.1. The structure of single-layer neural network.

식 (2.2)에서 w는 p-차원의 가중치 벡터이고 b는 편향(bias) 상수이다. 초평면 상의 점들은 $f(x) = 0$이다. 식 (2.2)의 분리초평면과 가장 가까이 위치하는 훈련용 자료와 분리초평면 간의 거리를 마진(margin)이라고 하고 이때, 분리초평면의 마진 상에 정확히 위치하는 점을 서포트 벡터(support vector)라고 한다. 서포트 벡터 머신은 마진을 최대로 하는 초평면을 찾는 것이다.

하지만 실제 데이터를 분석하고 예측함에 있어 모든 군집을 오차 없이 정확하게 분리할 수 없는 경우가 대부분이다. 오차를 허용하지 않고 모든 자료를 완벽하게 분류하게 된다면 마진이 매우 작아질 수 있으며 과대적합의 문제가 발생할 수 있다. 때문에 완벽한 분리에 초점을 맞추기보다 마진을 좀 더 넓히고 오분류율을 최소화하는 것을 목표로 하고, 분류 불가능한 데이터에 대해서는 벌점을 두는 방식으로 진행된다. 이때의 선형제약식은 다음과 같다.

$$\begin{aligned} & y_i f(x_i) > M(1-\epsilon_i), \\ & \epsilon_i > 0, \ \sum_i = 1^n \epsilon_i \leq C. \end{aligned} \tag{2.3}$$

여기서, $y_1, y_2, \ldots, y_n \in \{-1, 1\}$이고, 식 (2.3)은 표본이 오분류될 가능성을 허용하는 제약식이다. ϵ_i를 여유변수(slack variable)라고 하는데, 이는 오분류가 생기는 점에 적절한 페널티를 부여함으로써 비선형적으로 분리되는 모형을 완화시키는 선형제약 값이다. C는 비음 조율 모수(nonnegative tuning parameter)로, 목적함수를 구성하는 마진을 최대화하는 $||w||^2$와 오차항 $\sum_{i=1}^n \epsilon_i$ 사이에 존재하는 조정자이며 벌점 효과와 관련이 있다.

서포트 벡터 머신은 비선형결정경계(non-linear decision boundaries)를 가지는 문제도 해결할 수 있다. 이때의 서포트 벡터 머신의 핵심 아이디어는 원래의 p-차원의 공간을 커널함수(kernel function)를 이용해 고차원으로 매핑하여 선형적으로 분리가 가능하도록 하는 것이다. $K(x_i, x) = \Phi(x_i) \bullet \Phi(x)$를 커널함수라고 하며, 이는 매핑공간에서의 내적과 동등한 함수이다 (Schölkopf와 Smola, 2002).

또한 예를 들어 3개의 이상의 범주를 가지는 경우, k-번째 범주(+1로 코딩)와 이를 제외한 나머지 범주(−1로 코딩)를 비교하는 서포트 벡터 머신을 적합하여 분리초평면을 얻은 후, 검증용 자료를 이용하여 $f(x)$의 값이 가장 크게 나타나는 범주로 관측값을 분류한다.

2.3. 인공신경망

인공신경망의 구조는 크게 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)로 나뉘게 되며, 입력층과 은닉층 간의 노드(node)가 갖는 가중값에 따라 출력층의 값을 결정하게 된다. 이때, 은닉층은 활성 함수(activation function)로 구성되어 있으며 시스템이 복잡해질수록 층수가 늘어나고 출

력값에 대한 정확도가 향상되는 장점을 가진다. 자료가 입력되면 각 입력 자료에 대해 가중치를 부여하고 활성화 함수를 통해 출력값을 나타낸다. 다음의 Figure 2.1은 은닉층 없이 입력층이 직접 출력층에 연결되는 단층신경망(single-layer neural network)의 네트워크 구조이다.

Figure 2.1에서 z는 입력자료들의 가중합 나타낸다. 학습(learning 또는 training)을 거쳐 원하는 결과가 나오도록 오차가 작아지는 방향으로 가중치가 조정된다. 입력자료들의 가중합은 다음과 같은 식으로 나타낼 수 있다.

$$z = w'x + w_0,$$
$$\left(\text{또는 } z = w_0 + \sum_{j=1}^{p} w_j x_j \right).$$

여기서, $w = (w_1, w_2, \ldots, w_p)'$으로 가중치를, w_0는 편의를 나타내며, x는 p-차원의 입력벡터 $(x_1, x_2, \ldots, x_p)'$를 나타낸다. z값에 대해 활성함수가 적용되어 $y(x)$가 계산되는데, 가중치 w와 절편 w_0는 학습을 통해 오차제곱합이 최소가 되는 방향으로 갱신(update)된다. 다층신경망의 가중치는 학습과정에서 오차의 역전파(back-propagation) 알고리즘을 통해 갱신된다. 역전파 알고리즘이란 오차를 출력노드로부터 입력노드까지 역으로 전파하는 알고리즘으로 인공신경망을 학습시키기 위한 가장 기본적이고 일반적인 알고리즘이라고 할 수 있다.

2.4. 가중 k-인접이웃분류

k-인접이웃 분류는 새로운 데이터(설명변수 값)로부터 거리가 가까운 순서대로 k개의 과거자료 (설명변수 값)를 찾아서 그 중 가장 많은 수의 데이터가 속한 그룹의 항목을 할당하는 방법이다. 하지만 이 방법은 데이터 간의 거리의 정도를 생각하지 않고 가장 가까운 데이터 k개 안에서 다수결에 따라 그룹을 선택한다는 단점이 있다.

이를 해결하기 위해 가중 k-인접이웃분류(weighted k-nearest neighbor; Weighted k-NN)를 고려하게 되는데 이는 더 가까운 이웃일수록 더 먼 이웃보다 평균에 더 많이 기여하도록 거리에 대해 유사성 가중치를 주는 방법이다. 가중치는 역함수, 가우스 함수 등을 이용하여 부여한다.

가중 k-NN 분류를 수행할 때 거리함수를 고려해야한다. 새로운 데이터와 학습데이터 간의 거리를 기준으로 하여 이웃한 데이터를 찾게 되는데, 어떤 종류의 거리함수를 사용하는지에 따라 선택되는 이웃이 달라질 수 있다. 거리함수의 종류로는 가장 일반적으로 이용하고 있는 다음과 같은 유클리디안 거리 함수를 주로 사용한다.

$$d_E(x, y) = ||x - y||_2 = \sqrt{(x-y)^T(x-y)} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

2.5. 모형 평가 기준

기계학습에서 분류 (예측)모형에 대한 평가는 보통 전체 자료를 훈련용 자료와 검증용 자료로 나누어 훈련용 자료를 이용해 모형을 구축하고 구축된 모형을 검증용 자료에 적용하여 평가한다. 순서형 다범주 반응변수에 대한 모형평가 측도로는 정확도(accuracy)와 가중카파계수(weighted Cohen's kappa coefficient) 등이 사용된다. 순서형으로 이루어진 세 개의 그룹($k = 1, 2, 3$)를 갖는 분류모형의 정오분류표는 Table 2.1과 같다.

Table 2.1. The confusion matrix and weighting method

		실제집단		
		1	2	3
예측집단	1	$n_{11}(0)$	$n_{12}(1)$	$n_{13}(2)$
	2	$n_{21}(1)$	$n_{22}(0)$	$n_{23}(1)$
	3	$n_{31}(2)$	$n_{32}(1)$	$n_{33}(0)$

정확도는 전체 자료 중 실제로 올바르게 예측한 비율로, 전체 자료 수를 N이라 할 때 Table 2.1로부터 다음의 식을 갖는다.

$$\frac{n_{11} + n_{22} + n_{33}}{N}.$$

카파계수(Cohen's kappa coefficient)는 범주형 자료에 대한 일치도 측도(measure of agreement)로, 두 관찰자 간의 측정 범주 값에 대한 일치도를 측정하는 방법이다. 1은 완벽한 일치를 나타내고 0은 일치 결여를 나타낸다. 카파계수는 관찰자 간의 상대적으로 측정된 일치도(정오분류표의 정확도와 동일)와 관찰자에 의해 우연히 일치된 평가를 받을 비율(관측된 자료로부터 측정한 기대빈도로 이루어진 행렬에서의 정확도)로 이루어져 있다. 카파계수는 두 관찰자 간의 불일치를 고려하지만 이 불일치의 심각성까지는 고려하지 않는다. 순서형 반응변수의 경우, Table 2.1로부터 실제 '2'의 범주를 '1'의 범주로 분류(예측)하는 것보다 실제 '3'의 범주를 '1'의 범주로 분류 (예측)하는 것에 더 높은 가중치를 부여한다. 설명한 가중치는 Table 2.1에서의 각 셀의 괄호의 값과 같은 방식으로 부여한다. 가중카파계수의 공식은 다음과 같다.

$$K = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}}.$$

이때, i는 예측집단의 순서형 범주를 숫자로 부여한 값, j는 실제집단의 순서형 범주를 숫자로 부여한 값, k는 그룹의 수를 의미한다. w_{ij}는 가중치 행렬의 원소, x_{ij}는 관측된 행렬의 원소, m_{ij}는 관측된 자료로부터 측정한 기대빈도로 이루어진 행렬의 원소이다.

3. 실증분석

3.1. 분석자료

본 연구에서 사용한 자료는 충북대학교 수의과대학에서 실험한 세포독성 자료이다. 세포독성은 화학물질 등이 세포에 상해를 입히는 경우와 같은 특성을 말하는데 본 논문에서 사용된 세포독성 자료는 mEBT 방법으로 얻어졌다. mEST방법은 ECVAM에서 제안한 방법으로, 화학물질을 주입하였을 때 미분화된 쥐의 배아줄기세포 및 쥐의 분화된 섬유아세포의 생존력이 50%가 되는 시점과 mESC의 심근세포로의 분화가 50%가 되는 시점에서의 약물의 농도를 평가한다. 심근분화를 유도하기 위해서는 배상체 형성이 반드시 필요한데 이때 약물에 의해 배상체 크기가 감소하게 된다. 이러한 배상체 단계에서 발생독성을 평가하는 경우, 심근세포로의 분화 유도에 소요되는 시간과 노동력을 감소시킬 수 있으므로 mESC의 심근 형성을 쥐의 배상체의 단면적으로 대체하여 실험을 하게 되는데, 이를 mEBT방법이라고 한다. mESC 및 쥐의 분화된 섬유아세포의 생존력이 50%가 되는 시점에서의 세포독성을 평가하는 것은 기존의 mEST방법과 동일하지만 mESC의 심근세포로의 분화가 50%가 되는 시점이 아닌 mEB의 단면적이 50%가 되는 시점의 약물의 농도를 평가하는 것이 mEST방법과는 다른 점이라 할 수 있다.

Table 3.1. Class of data

	독성 분류	빈도
Class (약물의 독성 여부)	Non-toxic(무독성)	24
	Moderate toxic(중간 독성)	24
	Strong toxic(강한 독성)	20

Table 3.2. Explanation of independent variables

변수	설명
IC_{50}ESC	미분화된 쥐의 배아줄기세포에서 약물을 10일 동안 노출시킨 후의 세포 생존율이 절반으로 감소할 때의 약물의 농도
IC_{50}3T3	쥐의 섬유아세포에서 약물을 10일 동안 노출시킨 후의 세포 생존율이 절반으로 감소할 때의 약물의 농도
ID_{50}EB	쥐의 배상체에서 약물을 10일 동안 노출시킨 후의 면적이 절반으로 감소할 때의 약물의 농도

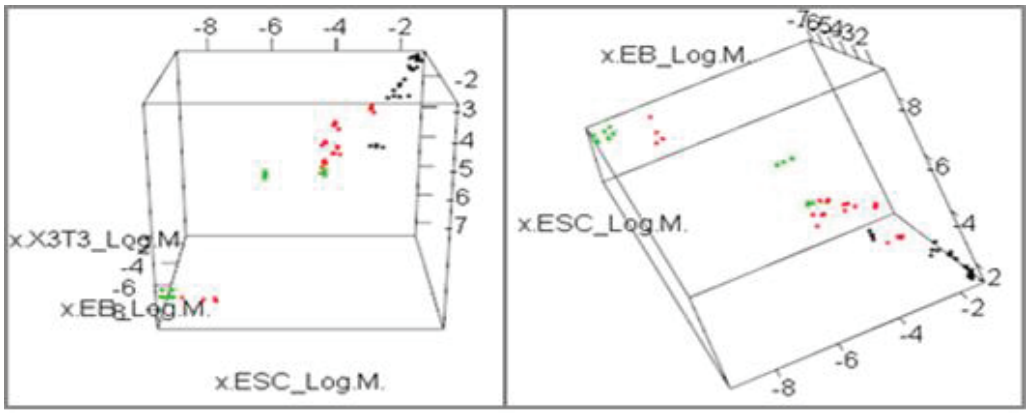

Figure 3.1. Scattering of drug concentration by three groups in the degree of toxicity.

총 15종의 화학물질을 쥐의 세포에 주입하여 총 68회의 실험을 실시하였으며 쥐의 각 세포 및 배상체에서의 측정값을 통해 약물의 독성을 예측한다. 독성은 무독성, 중간독성, 강한 독성인 세 단계로 분류하였다. 각 범주에 대한 자료의 개수는 무독성이 24개, 중간독성이 24개, 강한 독성이 20개로, 총 68개이다. Table 3.1에서는 자료의 분류 그룹을 Table 3.2에서는 자료의 독립변수를 나타내었다.

실험필드에서 해당 수치에 대하여 상용로그를 취하여 다루기 때문에 모든 독립변수에 상용로그를 취해 분석을 실시하였다. Figure 3.1은 각 변수들에 대한 약물의 농도를 3차원 그림으로 나타낸 결과이다. 그림을 각기 다른 두 방면에서 보았을 때를 나타낸 것이며, 검은색 점들은 무독성, 붉은색 점들은 중간독성, 연두색 점들은 강한 독성을 의미한다. 각 집단은 선형적 분리가 어려운 것을 파악할 수 있다.

3.2. 분류 알고리즘 분석

3.2.1. 이차판별분석 선형판별분석에서는 모집단에 대한 다변량 정규성, 그룹-내 공분산행렬의 동일성, 변수들 간의 낮은 다중공선성의 가정이 요구된다. 세 변수 간 상관분석을 실시한 결과, 모든 변수들 간의 상관관계가 0.9 이상으로 매우 높은 것을 확인할 수 있었다. 이에 분산팽창지수를 확인한 결과로는 모든 변수에서 10 이상의 값을 보였으므로 약한 다중공선성이 존재함을 확인하였다. 하지만

Table 3.3. The results of quadratic discriminant analysis

조각	정확도	가중카파계수
$i = 1$	0.941	0.922
$i = 2$	0.882	0.872
$i = 3$	0.941	0.931
$i = 4$	0.882	0.821
평균	0.912	0.887

Table 3.4. The results of support vector machine

조각	C	γ	정확도	가중카파계수
$i = 1$	5	1	0.941	0.922
$i = 2$	10	2	1.000	1.000
$i = 3$	10	2	1.000	1.000
$i = 4$	100	1	1.000	1.000
평균			0.985	0.980

mEBT방법을 사용하여 세포독성을 평가하는 해당 시험법은 세 변수를 모두 사용해야 학문적으로 의미가 있으므로 독립성을 가정하고 분석을 진행하였다. 더불어 Box-M 검정을 통한 등분산성 검정 결과, $\chi^2(6) = 136.215$로 오차의 등분산성이 만족되지 않는 것이 확인되었다. 따라서 그룹별로 개별 공분산 행렬을 가정하는 QDA를 수행하였다.

4-folds CV를 통해 모형 구축 및 평가를 하였으며, 각 조각별 분류 정확도와 가중카파계수는 Table 3.3에 제시하였다. QDA 수행 결과, 평균 분류 정확도가 0.912, 평균 가중카파계수가 0.887로 나타났다.

3.2.2. 서포트 벡터 머신 데이터를 선형으로 분리할 수 없는 경우에는 저차원의 공간을 커널 함수를 이용하여 고차원으로 맵핑한 후 선형적으로 분리가 가능하도록 한다. 본 연구에서는 가우시안 커널을 사용하였고, 가우시안 커널 함수식은 다음과 같다.

$$K(i,j) = \exp\left\{-\frac{||x_i - x_j||^2}{2\sigma^2}\right\}. \tag{3.1}$$

식 (3.1)에서 $\gamma = 1/2\sigma^2$로 정의되며 상수 C와 γ는 조율을 통하여 최적의 값으로 지정하였다. 4-folds CV를 수행한 후, 각 조각별 C와 γ, 이를 사용하여 얻은 분류 정확도와 가중카파계수는 다음의 Table 3.4에 나타내었다. SVM 수행 결과, 평균 정확도는 0.985, 가중카파계수는 0.980으로 나타났다.

3.2.3. 인공신경망 과적합을 방지하기 위해 은닉층은 하나로 제한하였고, 은닉 노드의 수는 입력 노드의 2/3개인 2개로 설정하였다. 활성함수는 다범주 분류에 적합한 softmax로 지정하였다. softmax는 0과 1 사이의 값으로 출력이 되도록 변환을 시켜주며 변환된 결과의 합계가 1이 되도록 한다. 해당함수는 다음과 같다.

$$P(y = j|X) = \frac{\exp(z_j)}{\sum_{k=1}^{K}\exp(z_k)}, \quad \text{for} \;\; j = 1, 2, 3,$$
$$z = w^T X + b.$$

오차함수는 분류의 목적으로 사용할 때 적합한 Cross-Entropy 함수로 지정하였다. 신경망에서는 과적합을 방지하기 위해 가중치가 클수록 페널티를 부과하게 되는데 가중치 행렬을 w라고 할 때 손실함수의

Table 3.5. The results of artificial neural network

조각	정확도	가중카파계수
$i = 1$	0.941	0.922
$i = 2$	1.000	1.000
$i = 3$	0.941	0.934
$i = 4$	0.824	0.773
평균	0.926	0.907

Table 3.6. The results of k-nearest neighbor

조각	정확도	가중카파계수
$i = 1$	0.941	0.922
$i = 2$	1.000	1.000
$i = 3$	1.000	1.000
$i = 4$	0.941	0.919
평균	0.971	0.960

Table 3.7. Comparison of the performance of four classification algorithms

분류 알고리즘	정확도	가중카파계수
QDA	0.912	0.887
SVM	0.985	0.980
ANN	0.926	0.907
k-NN	0.971	0.960

QDA = quadratic discriminant analysis; SVM = support vector machine; ANN = artificial neural network; k-NN = k-nearest neighbor.

결과에 $\lambda w^2/2$를 더해주어 정도를 조정하게 된다. 이때 λ를 0.001로 설정하였으며, 초깃값은 -0.1에서 0.1 사이의 값으로 랜덤하게 지정하였다.

다음의 Table 3.5는 4-folds CV를 수행한 후, 각 조각별 신경망 적합 과정과 분류 정확도, 가중카파계수를 나타낸 결과이다. ANN 수행 결과, 평균 정확도는 0.926, 평균 가중카파계수는 0.907로 나타났다.

3.2.4. 가중 k-인접이웃분류 가중 k-NN은 거리가 가까운 데이터에 가중치를 더 크게 부여하게 되는데 본 논문에서는 가중치 부여 방법은 다음과 같다.

$$K(d) = (1 - |d|) \cdot I(|d| \leq 1). \tag{3.2}$$

여기서, d는 데이터 간 거리를 나타낸다. 데이터 간 거리는 유클리디안 거리를 이용하여 측정하였다.

다음의 Table 3.6은 $k = 7$로 설정하고 4-folds CV를 수행한 후, 각 조각별 분류 정확도와 가중카파계수를 나타낸 결과이다. 여기서 k는 3, 5, 7, 9의 홀수 후보군 중 가장 예측력이 우수한 $k = 7$로 설정하였다. k-NN 수행 결과, 평균 정확도는 0.971, 평균 가중카파계수는 0.960으로 나타났다.

3.3. 분류 알고리즘 성능 비교

다음 Table 3.7는 네 모형의 성능을 비교하기 위해 분류 정확도와 가중카파계수를 정리한 표이다. 본 연구에서 사용한 자료는 이분산 자료이며 비선형 모형에 적합하므로 이차판별분석을 수행하여 분류 정확도 0.912, 가중카파계수 0.892의 결과를 얻었다. 실제로 공통 공분산 행렬을 가정하고 선형판별분석

을 실시한 결과 분류 정확도가 0.735로 이차판별분석이 훨씬 더 적합하다는 결론을 얻었다. 또한, Table 3.7를 보면 네 가지 알고리즘 모두 분류 정확도가 0.9 이상으로 예측력이 우수하였고, 가중카파계수 또한 0.8 이상으로 거의 완벽한 일치도를 보였다. SVM 모형의 정확도가 0.985, 가중카파계수 0.980으로 네 가지 분류 알고리즘 중 가장 좋은 성능을 나타낸 것을 알 수 있다. 따라서 최종 모형으로 SVM을 선택하고 전체 세포독성 자료를 훈련용 집합과 검증용 집합으로 각각 7 : 3으로 나누어 SVM의 성능을 평가한다.

최종적으로 전체자료의 70%인 훈련용 자료 49개를 이용해 SVM 모형을 적용하였다. 상수 C와 γ은 조율 작업을 통하여 각각 최적의 값인 10과 1로 지정하였고 해당 모형으로 검증용 자료 19개에 적용한 결과, 분류 정확도 0.947, 가중카파계수 0.927의 결과를 얻었다.

본 연구 자료에 있어서는 일반적으로 머신러닝의 기법이 우수하였으며 머신러닝 기법 중에서도 SVM 모형이 가장 우수하다는 결론을 얻었다.

4. 결론

세포독성 자료를 이용하여 네 가지 분류기법(이차판별분석, 서포트 벡터 머신, 인공신경망, 가중 -인접 이웃분류)의 성능을 비교하였다. 성능 비교의 측도로는 분류 정확도와 가중카파계수를 이용하였으며, 4-folds CV를 수행하여 각 네 개의 조각에서의 분류 정확도와 가중카파계수의 평균을 계산하여 해당 모형의 최종 측도로 사용하였다. 가장 바람직한 최종 분류 알고리즘을 선택한 후에는 전체 자료를 다시 한 번 훈련용 집합과 검증용 집합을 7 : 3의 비로 나누어 모형을 다시 평가하였다.

본 연구 자료를 이용해 도출한 예측 분류 결과를 살펴보면 전체적으로 머신러닝 기법이 우수한 것을 확인할 수 있었다. 최종 분류 알고리즘은 서포트 벡터 머신으로, 최종 분류 정확도 0.947, 가중카파계수 0.927의 성능을 보였다.

머신러닝의 경우, 분포에 대한 가정이 필요하지 않아 특별한 제약조건이 없지만 통계적 분석의 경우 여러 분포 가정들의 만족이 요구된다. 이는 본 연구 자료로 분석한 결과를 확인하였을 때 머신러닝 기법이 우수한 이유 중 하나라고 추측할 수 있다. 판별분석의 경우 여러 조건들을 충족시키지 못한 채 진행하였기 때문에 다른 방법들보다 예측율이 낮았다. 하지만 인공신경망과 비슷한 결과를 가져왔으므로 모든 가정들이 만족된다면 훨씬 더 좋은 분류 정확도를 보일 것으로 예상되며 머신러닝 기법을 수행한 결과와 별반 차이가 없을 것으로 기대된다.

References

Genschow, E., Scholz, G., Brown, N., *et al.* (2000). Development of prediction models for three in vitro embryotoxicity tests in an ECVAM validation study, *ALTA*, **13**, 51–66.

Kang, H. Y., Choi, Y. K., Jo, N. R., *et al.* (2017). Advanced developmental toxicity test method based on embryoid body's area, *Reproductive Toxicology*, **72**, 74–85.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*, Lon Don, The MIT Press.

Seiler, A., Visan. A., Buesen, R., Genschow, E., and Spielmann, H. (2004). Improvement of an in vitro stem cell assay for developmental toxicity, *Reproductive Toxicology*, **18**, 231–240.

Vapnik, V. N. (1998). *Statistical Learning Theory*, New York, Wiley.

Yu, S. G. (2010). *The Comparison of Classification Algorithms for Micro array data* (Graduate School master's thesis), Chungbuk National University.

세포독성 자료를 이용한 분류 알고리즘 성능 비교

윤여창[a] · 정의배[b] · 조나래[c] · 주수인[c] · 이성덕[c,1]

[a]우석대학교 정보보안학과, [b]충북대학교 수의학과, [c]충북대학교 정보통계학과

(2018년 5월 28일 접수, 2018년 6월 7일 수정, 2018년 6월 8일 채택)

요 약

최근 동물실험의 대체방법 중 하나로 쥐의 줄기세포 유래 배상체를 이용하여 독성을 시험하는 방법이 개발되었다. 이는 동물에 직접 약물을 주입하는 것이 아닌 배상체 세포에 약물을 투입하여 세포의 변화에 따른 측정값들을 얻는 방법이다. 본 연구에서는 다범주 세포독성 자료를 이용해 통계적 기법인 판별분석(discriminant analysis)과 머신러닝 기법인 서포트 벡터 머신(support vector machine), 인공신경망(artificial neural network), k-인접이웃분류(k-nearest neighbor)의 성능을 비교하였다. 알고리즘의 성능은 분류 정확도(accuracy)와 가중카파계수(weighted Cohen's kappa coefficient)로 비교하였다.

주요용어: 판별분석, 서포트 벡터 머신, 인공신경망, k-인접이웃분류, 세포독성

이 논문은 2017년도 식품의약품안전처의 연구개발비(17182MFDS487)로 수행된 연구임.

[1]교신저자: (28644) 충북 청주시 서원구 충대로 1, 충북대학교 정보통계학과. E-mail: sdlee@chungbuk.ac.kr