Regular paper | Journal of the KIECS. pp. 1353-1364, vol. 13, no. 6, Dec. 31 2018, t. 92, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2018.13.6.1353

불균형 데이터 환경에서 로지스틱 회귀모형을 이용한 *Cochlodinium polykrikoides* 적조 탐지 기법 연구

박수호* · 김흥민* · 김범규* · 황도현* · 엥흐자리갈 운자야* · 윤홍주**

Study on Detection Technique for *Cochlodinium polykrikoides* Red tide using Logistic Regression Model under Imbalanced Data

Su-Ho Bak* · Heung-Min Kim* · Bum-Kyu Kim* · Do-Hyun Hwang* ·
Unuzaya Enkhjargal* · Hong-Joo Yoon**

요 약

본 연구에서는 불균형 데이터 환경에서 기계학습 기법의 한 갈래인 로지스틱 회귀모형을 이용하여 인공위성 영상에서 *Cochlodinium polykrikoides* 적조 픽셀을 탐지하는 방법을 제안한다. 학습자료로 적조, 청수, 탁수 해역에서 추출된 수출광량 분광 프로파일을 활용하였다. 전체 데이터셋의 70%를 추출하여 모형 학습에 활용하였으며, 나머지 30%를 이용하여 모형의 분류 정확도를 평가하였다. 이 때, 청수와 탁수에 비해 자료 수가 상대적으로 적은 적조의 분광 프로파일에 백색 잡음을 추가하여 오버샘플링을 하여 불균형 데이터 문제를 해결하였다. 정확도 평가 결과 본 연구에서 제안하는 알고리즘은 약 94%의 분류 정확도를 보였다.

ABSTRACT

This study proposed a method to detect Cochlodinium polykrikoides red tide pixels in satellite images using a logistic regression model of machine learning technique under Imbalanced data. The spectral profiles extracted from red tide, clear water, and turbid water were used as training dataset. 70% of the entire data set was extracted and used for as model training, and the classification accuracy of the model was evaluated using the remaining 30%. At this time, the white noise was added to the spectral profile of the red tide, which has a relatively small number of data compared to the clear water and the turbid water, and over-sampling was performed to solve the unbalanced data problem. As a result of the accuracy evaluation, the proposed algorithm showed about 94% classification accuracy.

키워드

COMS/GOCI, Logistic Regression Model, Machine Learning, Ocean Color Remote Sensing, Red Tide
정지 궤도 해색 위성, 로지스틱 회귀 모델, 기계학습, 해수색 원격 탐사, 적조

* 부경대학교 지구환경시스템과학부
(shbak91@pukyong.ac.kr, funwarm@naver.com, bumkyu1005@nate.com, rupine725@hanmail.net, unuzaya.e@gmail.com)
** 교신저자 : 부경대학교 지구환경시스템과학부
· 접 수 일 : 2018. 11. 09
· 수정완료일 : 2018. 11. 27
· 게재확정일 : 2018. 12. 15
· Received : Nov. 09, 2018, Revised : Nov. 27, 2018, Accepted : Dec. 15, 2018
· Corresponding Author : Hong-Joo Yoon
Divsion of Earth Environmental System Science Major of Spatial Information Engineering, Pukyong National University,
Email : yoonhj@pknu.ac.kr

Ⅰ. 서 론

적조현상은 식물 플랑크톤이 특정 환경조건에서 대량으로 증식하여 해수면이 변색되는 현상이다. 그러나 근래에 해수면을 변색시키지 않는 저밀도의 식물플랑크톤이 다른 생물에게 피해를 입히는 현상들이 보고됨에 따라 해양에서 식물 플랑크톤 대량증식이 생물에게 물리적 피해를 야기하는 현상을 유해적조(HAB; Harmful Algal Bloom)라 구분하여 정의하고 있다[1].

우리나라의 경우 1980년대까지는 규조류(Diatoms)에 의한 적조가 남해연안 일부 해역에서 일시적으로 발생하였으나[2], 1990년대에 들어서는 와편모조류(Dinoflagellates)에 의한 적조발생 비율이 급격히 증가하여 적조발생 양상이 변화하고 있는 추세이다. 이들 와편모조류들은 대부분 유해적조를 일으키는 종으로 근래에 적조발생이 이슈가 되는 것은 이러한 추세 때문이다.

와편모조류는 규조류와 달리 두 개의 편모를 가지고 있으며, 이 편모를 이용하여 스스로 이동할 수 있다. 운동능력으로 인해 한번 대발생을 일으키면 저층의 높은 영양조건에서 지속적으로 영양분을 공급받아 장기간, 넓은 해역에서 적조현상을 지속시킬 수 있다[3]. 일반적인 규조류가 일주일 가량 적조현상을 지속시키는데 비해 와편모조류는 수 주일에서 수개월까지 지속시킬 수 있다. 이러한 와편모조류 중 특히 우리나라에서는 *Cochlodinium polykrikoides*(이후 *C. polykrikoides*)가 최근 20여년 간 매년 여름과 가을철(7~10월)에 발생하여 수산업에 막대한 피해를 주고 있으며, 특히 1995년에는 약 750억원 규모의 경제적 손실을 가져온 종이다. *C. polykrikoides*는 무독성이지만, 점액질을 생성하여 대발생을 일으켰을 경우 어류의 아가미에 부착하여 질식사시킴으로써 양식업에 큰 경제적 손실을 입하고 있다. 특히 2013년의 경우 동해안까지 확산되어 우리나라 연안 수산업에 막대한 손실을 입혔던 전례도 가지고 있다.

*C. polykrikoides*에 대한 생리 생태학적 연구가 지속되고는 있으나 아직 적조발생 원인이나 과정이 명확히 밝혀지지 않아 발생을 사전에 예측하여 대비하는 것이 어렵다[4]. 이러한 이유로 사전예측보다는 조기탐지 및 지속적인 모니터링을 통한 신속대응이 효과적이다[4, 19]. 그러나 기존의 적조 모니터링 방법은 규조류에 의한 적조를 위한 것으로 주로 선박이나 항공기를 이용해왔다. 이러한 방법은 인력, 비용, 시간적 측면에서 비효율적일 뿐만 아니라 광범위한 해역에 발생하는 와편모조류에 의한 적조를 모니터링하는데 제한사항이 많으므로 인공위성을 이용한 원격탐사 도입이 필요한 상황이다[5].

인공위성 원격탐사를 이용할 경우 현장 조사 인력 없이 넓은 해역을 탐지 및 모니터링 할 수 있으며, 시각적인 결과물을 얻을 수 있다는 장점이 있다[6].

이러한 원격탐사를 활용한 적조탐지 기법들은 초기에는 인공위성 영상 기반으로 산출된 Chlorophyll-a 농도를 이용하여 시도되었다[7-10]. 이 방법들은 적조가 발생하지 않은 일정기간의 Chlorophyll-a 농도 평균값을 기준으로 anomaly를 계산하여 특이값을 가지는 픽셀을 추출하는 기법이다. 그러나 적조를 유발시키는 종에 따라 Chlorophyll-a 농도를 높이는 정도가 다르며, 단순히 식물 플랑크톤의 증식으로 인한 Chlorophyll-a 농도 변화를 이용하게 될 경우 대발생을 일으킨 식물플랑크톤이 유해성인지 무해성인지 구분하지 못한다는 단점을 가지고 있다[6]. 또한 대부분의 해색위성 산출물에 적용되는 Chlorophyll-a 산출 알고리즘들은 청색 파장의 밴드와 녹색 파장의 밴드 간의 반사도 차이를 이용한 것으로 해수의 색이 식물플랑크톤에 의해서만 결정되는 맑은 해수(Case-1)에서는 신뢰할 수 있으나, 용존 유기물의 농도가 높은 우리나라 연안해역(Case-2)에서는 그 정확도가 낮다는 단점이 있다[11-13]. 우리나라 연안에서 발생한 적조는 용존 유기물 및 부유물질의 영향으로 청색 파장의 가시광선에서 흡광이 증가하여 위성에서 추정된 Chlorophyll-a 농도가 과대추정되며, 이로인해 고농도 용존 유기물 해역이 고농도 식물플랑크톤 해역으로 탐지될 수 있다[13].

최근에는 Chlorophyll-a 농도 기반 탐지기법의 단점을 개선하고자 해수의 분광학적 특성을 이용하여 시도하고 있다[4, 11, 14, 15]. 해수의 분광학적 특성은 외부의 광학적 환경에 관계없이 일정한 값을 가지는 흡광계수(Absorption coefficient; a), 산란계수(Scattering coefficients; b), 감쇄계수(Attenuation coefficient; c), 수출광량(Water leaving radiance; Lw), 원격반사도(Remote sensing reflectance; R_{rs}) 등이 있다[6]. 이러한 광특성은 물질의 종류에 따라

서로 다른 특성을 나타내므로 이를 역이용하여 해수 중의 물질 종류와 양을 추정할 수 있다[5, 16, 17]. 그러나 이 방법은 배경이 되는 해양의 가시광선 스펙트럼과 적조의 스펙트럼 형태 차이를 이용한 것으로 탐지 대상이 되는 식물 플랑크톤과 배경해수의 광학적 특성을 충분히 이해하고 있다는 가정 하에 제한적으로 활용될 수 있다. 또한 용존 유기물 및 부유물질의 농도가 높은 해수(Case-2)의 광학적 특성은 식물 플랑크톤 외의 다른 물질(용존 유기물, 부유물질)의 농도에 따라서 크게 변화하기 때문에 이들의 농도 변화에 따라 매우 다양한 배경해수색이 만들어질 수 있다. 이로 인하여 충분히 많은 상황에서 얻어진 탐지 대상의 스펙트럼을 얻어낸다 하더라도 배경해수의 본래 스펙트럼 형태에 따라 변색 패턴이 상이하기 때문에 모든 상황에 대처 가능한 탐지 식을 생산하는 것은 어렵다.

이러한 복잡한 배경해수에서 발생한 적조로 인한 변색을 모델링하기 위해서는 해수의 변색과정에 대한 높은 수준의 인과관계 이해가 필요하다. 그러나 Case-2 해수는 그 종류가 매우 방대하여 Case-2에 포함된 각기 다른 배경색이 만들어내는 색에 대한 인과관계를 규명하기는 어렵다. 이러한 문제를 해결하기 위해 최근 기계학습 기반 적조 탐지 알고리즘 개발 연구가 이루어지고 있다[20, 21]. 그러나 기계학습 기반 분류 알고리즘을 개발할 경우 학습데이터셋의 불균형 데이터 문제가 발생할 수 있다. 불균형 데이터는 특정 클래스의 레이블 수가 다른 클래스에 비해 현저하게 적거나 많은 경우를 의미하며 이는 기계학습 알고리즘의 성능을 저하시키는 요인으로 작용할 수 있다[22]. 이는 일반적인 기계학습 알고리즘들의 경우 학습데이터셋이 클래스별로 비슷한 비율로 구성되어 있다는 가정하에 학습을 진행하기 때문이다. 그러나 실세계의 많은 데이터들이 불균형 데이터 문제를 지니고 있으며, 이러한 경우 소수 레이블을 가진 클래스에 속한 데이터들은 다수 레이블을 가진 데이터보다 오분류될 가능성이 높아진다. 적조 현상 또한 연중 수일에서 수십 일 정도로 발생하며, 발생 위치 또한 한정적이기 때문에 학습데이터셋 확보가 어렵다. 따라서 기계학습 기반의 적조 탐지 알고리즘 개발 시 불균형 데이터 문제가 발생하게 된다.

본 연구에서는 이러한 기계학습 기반 적조탐지 알고리즘 개발 시 발생하게 되는 불균형 데이터 문제를 해결하기 위해 로지스틱 회귀모형 기반의 3단계 필터링 적조 탐지 알고리즘을 제안하고자 한다.

II. 자료 및 방법

2.1 데이터셋

본 연구에서는 해수의 유형을 청수, 탁수, 적조의 3가지로 분류하였다. 기계학습 모형을 학습시키기 위해 해수의 유형별로 가시광선 및 근적외선 영역 수출광량 분광 프로파일을 확보하여 활용하였다. 분광 프로파일은 해양위성센터에서 제공하는 GOCI Level 1B 자료를 통해 생산하였다. 또한 적조해역에서의 분광프로파일을 획득하기 위해 2013년부터 2015년까지의 국립수산과학원 적조속보 자료를 활용하였다. 2018년 적조의 경우 GPS를 활용한 현장관측을 통해 위치정보를 획득하였다.

GOCI(: Geostationary Ocean Color Imager)는 천리안 위성의 해양탑재체로 공간해상도는 500m이며, 가시광선 영역의 6개(412, 443, 490, 555, 660, 680 nm) 채널과 근적외 영역의 2개(745, 865nm) 채널을 가지고 있다(표 1). Level 1B 영상은 방사보정과 기하보정이 된 자료이며, 이를 GDPS(GOCI Data Processing System)를 이용하여 대기보정 하였다. 대기보정 후 산출되는 Level 2 자료 중 정규화 수출광량(Normalized Water-leaving radiance; $nLw(\lambda)$)을 사용하였다.

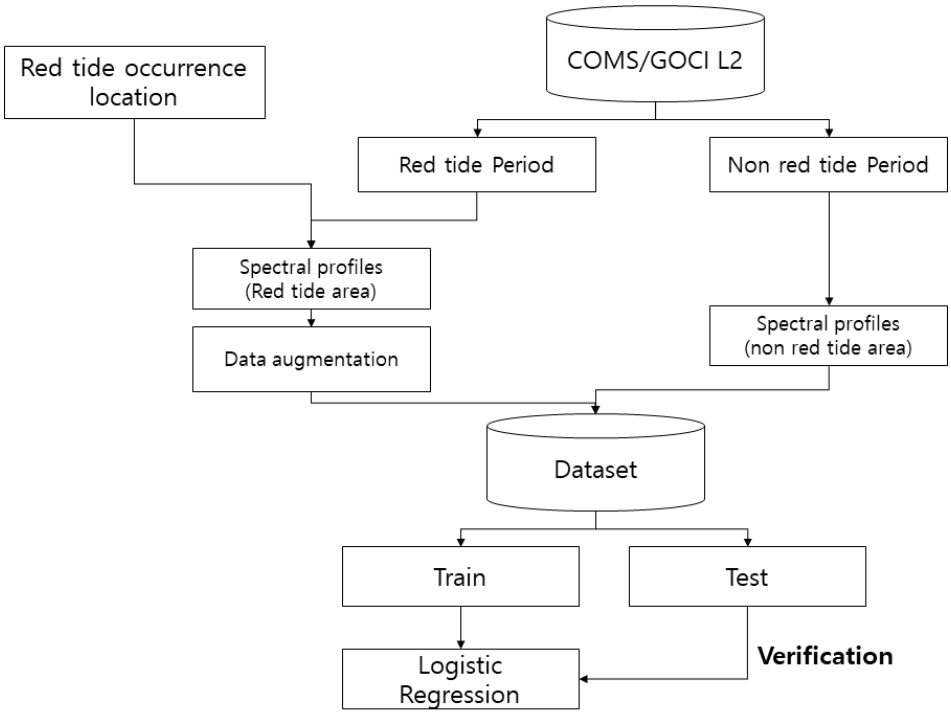

그림 1. 연구 흐름도
Fig. 1 Flow chart of research

표 1. GOCI 밴드 구성
Table 1. Band Composition of GOCI

Band	Centroid Wavelength (nm)	Bandwidth (nm)
1	412	20
2	443	20
3	490	20
4	555	20
5	660	20
6	680	10
7	745	20
8	865	40

국립수산과학원에서 제공하는 적조속보 자료는 적조발생 시 매일 1회 또는 필요에 따라 그 이상 제공되며, 적조발생 해역의 위치와 해당 위치에서의 원인생물, 생물밀도, 수온 등에 대한 정보를 포함하고 있다. 본 연구에서는 과거 적조발생 위치 정보를 수집하기 위해 적조속보 자료에 포함된 적조발생 해역도를 활용하였다. 적조발생해역도는 이미지 형태 자료로 정확한 위치정보를 포함하고 있지 않기 때문에 지오레퍼런싱(Georeferencing)을 통해 공간자료화하여 적조발생 해역의 위치정보를 추출하였다.

지오레퍼런싱은 항공사진 또는 이미지 형태의 지도의 X, Y좌표와 실세계 좌표를 일치시키는 방법이다. 사진 또는 지도 상의 기준점에 실제 좌표를 입력함으로써 이루어지며, 그 외 모든 좌표는 이 기준점에 대한 상대적인 좌표값으로 입력된다.

최종적으로 확보된 분광 프로파일 데이터셋은 약 600,000개의 레이블로 구성되어 있으며, 이 중 청수와 탁수가 각각 300,000개로 적조의 분광 프로파일은 1000여 건에 불과하다. 이는 적조현상이 연중 특정 시기에 특정 해역에서만 관측가능 한 이벤트로 청수와 탁수에 비해 분광프로파일을 얻기 힘들기 때문이

다. 본 연구에서 활용한 데이터셋은 전형적인 불균형 클래스 데이터셋으로 이 문제를 해결하지 않을 경우 기계학습 알고리즘의 성능이 저하될 수 있다.

이러한 불균형 데이터 환경에서 기계학습 모형을 학습시키기 위해 Bak et al.(2018)은 적조 분광 프로파일의 레이블 수에 나머지 클래스의 레이블 수를 맞추는 언더샘플링을 적용시켰다. 그러나 이러한 언더샘플링을 하게 될 경우 그 유형이 매우 다양한 탁수(Case-2)의 일부 사례만 샘플링되는 문제가 발생할 수 있다. 이 경우 데이터셋이 포함되지 못한 많은 유형의 탁수들과 적조 픽셀을 혼동하여 탁도가 높은 연안 해역에서 오탐지율을 높일 수 있다.

따라서 본 연구에서는 탁수 클래스의 분광 프로파일을 최대한 학습과정에서 반영하기 위해 적조 분광 프로파일을 오버샘플링하였다. 다른 클래스에 비해 상대적으로 레이블 수가 적은 적조의 분광 프로파일을 70:30으로 분리시켜 전체 데이터셋의 70%를 학습데이터로 사용하였고 나머지 30%는 검증용 데이터로 사용하였다. 이중 학습데이터에 인위적인 백색 잡음을 추가하여 그 수를 다른 클래스와 동일한 300,000개로 증폭시켰고, 이를 이용하여 학습데이터셋의 클래스 불균형 문제를 해결하였다(그림 1).

2.2 로지스틱 회귀모형

원격탐사 산출물 개발 시 탐지 목표물과 탐지에 활용하는 센서의 밴드들 사이의 관계를 모델링하는 작업이 필요하다. 이때, 탐지 목표물을 설명하기 위해 어떤 밴드 조합을 선택할 것인가에 대한 문제가 발생하게 된다. 특징 선택(Feature Selection)이라 불리는 이 과정은 최종적인 탐지 모델의 성능에 큰 영향을 미치므로 매우 중요하다. 적조 산출물을 개발하는데 있어서 선행연구들은 적조를 탐지하기 위해 적조의 분광프로파일을 시각적으로 판독하여 특징선택을 수행해왔다. 이러한 시각적인 판독으로도 일반해수와 적조해수 사이에 매우 큰 차이를 보이는 밴드 조합을 만들어낼 수 있었으나, 이는 연구자의 관찰력과 경험에 의존적인 것으로 연구자가 발견하지 못한 특징은 모델링에 활용될 수 없다는 단점이 있다. 반면 기계학습 기법을 적용할 경우 축적된 데이터에 의존한 특징선택을 수행하기 때문에 경험적 방법에 비해 객관성을 확보할 수 있으며, 연구자가 시각적으로 판독하지 못한 특징들이 모델에 반영되어 정확도 향상에 기여할 수 있다. 따라서 본 연구에서는 기계학습 기법 중 로지스틱 회귀모형(Logistic Regression Model)을 활용하여 *C. polykrikoides* 적조 탐지를 시도하였다.

로지스틱 회귀 모형은 독립변수와 종속변수 사이의 관계를 함수로 나타내는 통계적 모형으로 일반적인 회귀모형과 유사하다. 일반적인 회귀모형과 같이 설명변수 간의 선형 결합을 통해 반응 변수를 설명하지만 범주형 자료가 입력값으로 사용되었을 때, 그 결과가 특정 클래스에 속하게 될 확률로 주어진다는 데서 차이를 보인다[21]. 또한 일반적인 회귀모형에서는 반응변수의 범위가 $-\infty$에서 $+\infty$의 값을 가지는데 비해 로지스틱 회귀모형의 경우 로지스틱 함수(Logistic Function)를 통해 종속변수의 값이 변환되어 0에서 1의 값을 가진다. 로지스틱 회귀모형의 회귀계수는 일반적인 선형회귀계수와 같이 반응 변수와 설명변수들 사이의 관계를 설명하는데 사용된다[17, 21].

로지스틱 회귀모형을 포함한 모든 회귀모형은 설명변수의 선택 방법에 따라 모형의 성능이 영향을 받을 수 있다. 반응 변수에 영향을 줄 수 있는 선택가능한 모든 설명변수를 이용하여 모형을 생성할 경우 데이터를 획득하고 분석하며 관리하는데 많은 노력과 시간이 요구될 뿐만 아니라 오히려 모형이 종속변수를 설명하는 능력을 저해시킬 수 있다[21]. 따라서 반응변수를 설명할 수 있는 최소한의 설명변수 조합을 찾아내어 사용하는 것이 중요하다[21]. 회귀모형의 변수조합을 찾아내는 수학적 방법에는 일반적으로 후진소거법, 전진선택법, 단계적 선택법이 있다.

후진소거법은 모든 설명변수를 포함한 모형을 생성한 후 일정한 유의수준을 만족시키지 못하는 변수를 찾아내어 순차적으로 제거하는 방법이다. 그러나 모형에서 한번 제거된 설명변수는 다시 모형에 포함될 수 없기 때문에 소거를 수행하는 과정에서 더 좋은 설명변수 조합이 있더라도 선택할 수 없다[21].

전진선택법은 1개의 설명변수만을 이용하여 모형을 생성한 후 설명력이 상대적으로 우수한 변수를 모형에 추가하는 방법이다. 후진소거법과 함께 많이 사용되는 선택법이나 모형에 한번 포함시킨 설명변수는 도중에 탈락시킬 수 없기 때문에 유의하지 못한 변수가 포함될 가능성이 있다[21].

단계적 선택법은 후진소거법과 전진선택법의 장점

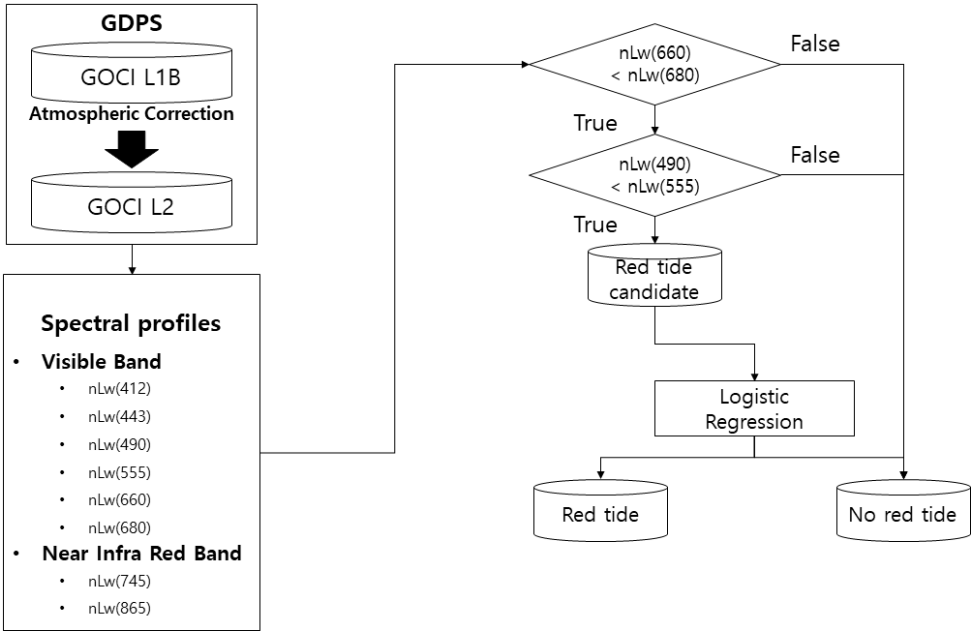

그림 2. 로지스틱 회귀모형을 활용한 적조탐지 과정
Fig. 2 Process of Red tide detection using Logistic regression model

을 취한 방법으로 모형에 포함된 변수도 선택과정에서 언제든지 제거될 수 있다. 단, 모든 변수조합을 모의해야하므로 후진소거법과 전진선택법에 비해 변수의 수 증가에 따른 연산시간 증가폭이 크다는 단점이 있다.

본 연구에서는 단계적 선택법을 통해 로지스틱 회귀모형의 최적 변수조합을 선택하였으며, AIC(Akaike Information Criterion)을 최소화하는 모형을 최적모형으로 정의하였다.

AIC는 변수선택법을 적용할 때, 이전 단계에서 선택된 변수조합에 비해 새롭게 생성된 변수조합이 얼마나 효과적인지 판단하는 기준으로, 회귀모형의 복잡한 정도에 벌점을 부여하여 최적화시키는 방법이다. AIC는 k개의 변수를 설명변수로 가지는 회귀모형 M_k에 관하여 식 (1)과 같이 정의된다[18, 21].

$$\mathrm{AIC} = \frac{2k}{n} - 2\sum_{i=1}^{n}\frac{L(y_i, x_i^T\hat{\beta})}{n} \quad (1)$$

이 때, $\sum_{i=1}^{n} L(yi, x_i^T\hat{\beta})$는 로그 우도함수이며, $\hat{\beta}$는 β에 대한 최대우도 추정량을 나타낸다. AIC는 회귀모형이 채택한 설명변수의 수 k가 증가함에 따라 비례하여 값이 커지므로, 상대적으로 더 많은 설명변수를 선택하게 될 경우 모델의 적합도를 낮게 평가하게 된다.

2.3 로지스틱 회귀모형을 활용한 적조탐지 알고리즘

본 연구에서 제안하는 적조탐지 알고리즘은 총 3단계의 과정을 포함하고 있다(그림 2). 이 중 첫 번째와 두 번째 단계를 통해 적조 후보 픽셀을 판별하며, 세 번째 단계에서 미리 학습시켜둔 로지스틱 회귀모형을 이용하여 적조 픽셀을 탐지하게 된다.

먼저, $660nm$와 $680nm$에서의 수출광량을 비교하며, 그 다음 단계에서 $490nm$와 $555nm$에서의 수출광량을 비교하여 적조 후보 픽셀을 분리시킨다. C. polykrikoides의 생물밀도가 증가하게 되면 세포 내에 존재하는 광합성 색소인 Chlorophyll-a와 카로티노이

드계 보조 색소의 양도 함께 증가하게 된다. Chlorophyll-a는 청색 파장(약 440~450nm)과 적색 파장(약 650~670nm)을 흡수하며, 680nm~690nm 파장대에서 형광신호를 방출한다. 또한 보조 색소인 카로티노이드계 색소로 인해 450~500nm의 청색 파장이 흡수된다[10-12]. 이로 인해 적조가 발생하게 되면 색소의 증가로 490nm와 660nm 파장에서의 흡광이 일어나 상대적으로 555nm와 680nm 파장에 비해 낮은 수출광량 값을 보이게 된다. 결과적으로 2단계 필터링 과정을 거치게 되면, 일반적인 청수와 탁수 픽셀이 제거된다. 이후 고밀도 식물플랑크톤과 유사한 분광 프로파일을 보이는 픽셀들을 로지스틱 회귀모형에 입력시켜 최종적으로 *C. polykrikoides* 적조 픽셀만 탐지하게 된다(그림 2).

2.4 정확도 평가

본 연구에서는 분류 정확도를 평가하기 위해 혼동행렬(Confusion Matrix, 표 2)을 이용하였다. 혼동행렬은 기계학습 모형의 정확도를 평가하기 위해 사용되는 평가측도 중 하나로 다수의 클래스를 분류하는 모형의 정확도를 정량적으로 평가하는데 많이 사용된다.

표 2. 혼동행렬
Table 2. Confusion Matrix

		Prediction	
		True	False
Reference	True	TP	FN
	False	FP	TN

※ TP : True Positive / FP : False Positive
FN : False Negative / TN : True Negative

이 때, 모형의 예측 또는 분류 정확도(Accuracy)는 식 (2)와 같이 정의된다.

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN} \quad (2)$$

TP와 TN은 모델이 실제 클래스로 정확하게 분류한 사례 수이며, FN과 FP는 모델이 클래스를 혼동한 사례의 수이다. 따라서 정확도는 모델이 검증 데이터셋을 분류한 결과 중 정확하게 맞춘 사례의 비율을 뜻하게 된다.

기계학습 모형의 경우 일반적으로 데이터셋을 특정 비율로 나누어 그 일부를 학습용 데이터로 사용하며, 나머지를 검증용 데이터로 사용한다. 이 때, 데이터셋에 불균형 클래스 문제가 있을 경우, 검증용 데이터에도 클래스 간 불균형이 발생하게 된다. 불균형 데이터 문제가 있는 상황에서 혼동행렬을 이용하여 정확도를 평가하면 소수 클래스에 대한 혼동사례가 저평가되어 정확한 정확도 평가가 힘들어진다. 따라서 본 연구에서는 검증용 데이터 선별 시 클래스 간 샘플링 수를 일치시켜 검증에 활용하였다.

III. 결과 및 토의

3.1 모형 생성 결과

로지스틱 회귀모형을 생성하기 위해 단계적 선택법을 적용한 결과 선택된 설명변수는 $nLw(412)$, $nLw(443)$, $nLw(490)$, $nLw(555)$, $nLw(660)$, $nLw(680)$, $nLw(745)$로 총 7개 파장의 정규화 수출광량이었다. 해수에서 식물플랑크톤의 개체 수 증가로 Chlorophyll-a 농도가 증가할수록 440~450nm와 670~680nm 파장대에서 흡수가 일어나며, 560~570nm 파장대에서는 반사 670~680nm 파장대에서는 형광이 일어난다[10-12]. 이로 인해 적조발생한 해역에서는 높은 Chlorophyll-a 농도로 인해 555nm와 680nm의 두 파장에서 피크(Peak)를 보인다[10]. 변수 선택법을 적용하여 얻어진 7개 파장 중 $nLw(490)$, $nLw(555)$, $nLw(680)$, $nLw(745)$는 이러한 적조발생 해역에서 관찰되는 피크와 관련된 것들로 선행연구들이 제시한 알고리즘 개발에도 많이 활용되었던 파장들이다[4, 10, 12, 13].

3.2 정확도 평가 결과

로지스틱 회귀모형을 활용하여 적조 탐지를 시도했던 선행연구[21]의 결과와 본 연구에서 제안하는 알고리즘의 결과를 비교해보았다.

검증용 데이터셋을 이용하여 두 알고리즘의 성능을 평가한 결과 선행연구의 알고리즘은 약 96%, 본 연구에서 제안하는 알고리즘은 약 94%의 분류 정확도를 보였다(표 3과 4).

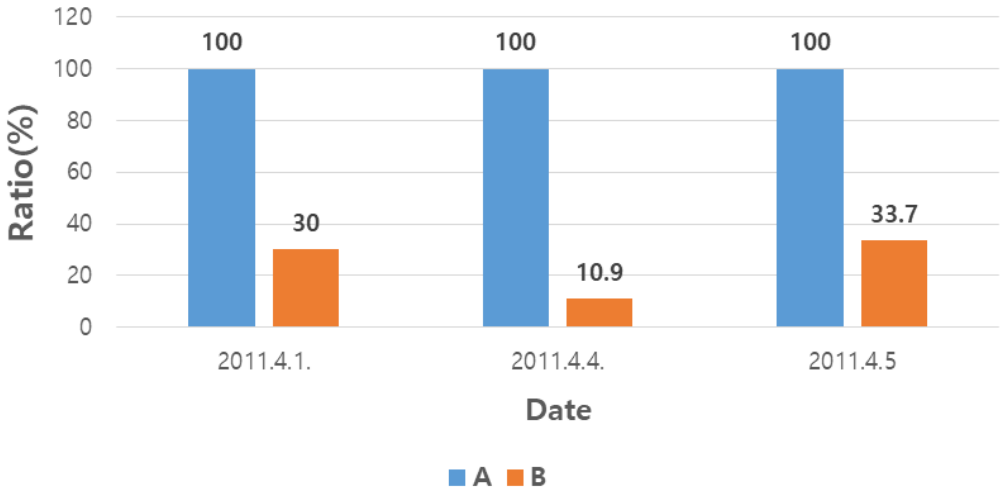

그림 3. 적조 미발생 기간에 적조로 탐지된 적조 픽셀 수의 상대적 차이(A : Bak et al., 2018, B : 제안하는 알고리즘).
Fig. 3 The relative difference in the number of red tide pixels detected as red tides during the no red tide occurrence period(A : Bak et al., 2018, B : Proposed Algorithm).

제안하는 알고리즘은 선행연구에 비해 탁수 픽셀을 적조 픽셀로 오탐지하는 능력은 개선되었으나, 실제 적조 픽셀을 탁수로 오탐지하는 사례는 오히려 선행연구에 비해 많았다. 그러나 전반적인 탐지 능력은 두 알고리즘 간에 큰 차이가 없었다.

표 3. 언더샘플링을 적용한 로지스틱 회귀모형[21]의 혼동행렬
Table 3. Confusion Matrix of Logistic Regression Model applied under-sampling[21]

		Reference		
		R	C	T
Detection	R	463	0	27
	C	0	478	0
	T	23	8	459

※ R: Red tide / T : Turbid water / C : Clear water
※ Total Accuracy≒0.96

표 4. 제안하는 알고리즘의 혼동행렬
Table 4. Confusion Matrix of Proposed algorithm

		Reference		
		R	C	T
Detection	R	412	0	9
	C	0	486	0
	T	74	0	477

※ R: Red tide / T : Turbid water / C : Clear water
※ Total Accuracy≒0.94

그러나 적조가 발생하지 않은 영상을 입력값으로 하여 두 알고리즘의 결과를 비교하였을 때, 본 연구에서 제안하는 알고리즘은 선행연구 대비 약 20% 수준의 낮은 오탐지율을 보였다(그림 3). 선행연구가 적조 미발생 기간 영상에서 적조로 탐지한 픽셀은 대부분 탁수 픽셀로 영상에서 탁수 해역 면적이 넓은 2011년 4월 5일 영상에서 가장 많은 적조 픽셀을 탐지하였다. 이는 학습데이터셋을 언더샘플링을 통해 구성하는 과정에서 탁수에 포함되는 많은 유형의 해수 분광 프로파일이 유실되어 상대적으로 적조의 분광 프로파일과 유사한 탁수 픽셀이 적조로 분류된 결과로 판단된다. 반면 본 연구에서 제안하는 알고리즘의 경우 선행연

구에 비해 1000배 이상의 탁수 분광 프로파일이 학습 데이터셋에 포함되어 더 많은 유형의 탁수 분광 프로파일을 적조와 구분시켜 학습시킬 수 있었다.

Ⅳ. 결 론

본 연구에서는 로지스틱 회귀모형을 활용하여 위성영상에서 적조 픽셀과 일반해수 픽셀을 분류하였다. 제안하는 알고리즘의 성능을 평가해본 결과 선행연구에서 탐지한 결과와 전체 정확도에서는 유사한 성능을 보였다. 그러나 적조가 발생하지 않는 기간에 촬영된 영상을 입력값으로 하였을 때는 선행연구에 비해 탁수 픽셀을 적조 픽셀로 오탐지하는 사례가 약 80% 가량 감소한 것을 확인 할 수 있었다.

이는 적조 탐지에 있어서 불균형 데이터 문제를 해결하기 위한 방법으로 언더샘플링 방식에 비해 백색잡음을 추가하는 오버샘플링 방식이 더 적합함을 의미한다.

적조 현상 또한 연중 수일에서 수십 일 정도로 발생하며, 발생 위치 또한 한정적이기 때문에 학습데이터셋 확보가 어렵다. 뿐만 아니라 원격탐사 대상이 되는 많은 자연현상들은 평상 시에 자주 발생하지 않는 특이한 현상인 경우가 많아 적조 현상과 같은 문제가 발생하기 쉽다. 따라서 기계학습 기법을 활용한 원격탐사 산출물 개발 시 많은 경우 불균형 클래스 문제에 직면할 수 있다. 따라서 본 연구의 결과는 적조 이외의 다른 원격탐사 산출물 개발에도 활용될 수 있을 것으로 예상된다.

감사의 글

본 논문은 2018년 해양수산부 재원으로 한국해양과학기술진흥원의 지원을 받아 수행된 연구임 (“적조피해 최소화를 위한 적조탐지·예측 시스템 구축 및 실증화”, PM60650)

References

[1] D. Anderson, P. Anderson, V. Bricelj, J. Cullen, and J. Rensel, *Monitoring and Management Strategies for Harmful Algal Blooms in Coastal Waters.* Paris: Intergovernmental Oceanographic Commission Technical Series, 2001.

[2] H. Kim, *Harmful Algal Blooms in the Sea.* Busan: Dasom, 2005.

[3] Y. Yoon, *Sea rebellion, Red tide.* Paju: Jipmoondang, 2012.

[4] Y. Kim, Y. Byun, Y. Huh, and Y. Yu, “Detection of Cochlodinium polykrikoides Red Tide Using MODIS Level 2 Data in Coastal Waters,” *Korean Society of Civil Engineers J. of Civil Engineering*, vol. 27, no. 4D, 2007, pp. 535-540.

[5] Y. Ahn, J. Moon, W. Seo, and H. Yoon, “Inherent Optical Properties of Red Tide Algal for Ocean Color Remote Sensing Application,” *J. of the Korean Society for Marine Environmental Engineering*, vol. 12, no. 1, 2009, pp. 47-54.

[6] S. Bak, H. Kim, B. Kim, D. Hwang, E. Unuzaya and H. Yoon, “Study on Detection Technique for Cochlodinium polykrikoides Red tide using Logistic Regression Model and Decision Tree Model,” *J. of the Korean institute of Electronic Communication Sciences*, vol. 13, no. 4, 2018, pp. 777-786.

[6] R. Stumpf, M. Culver, P. Tester, M. Tomlinson, G. Kirkpatrick, B. Pederson, E. Truby, V. Ransibrahmanakul, and M. Soracco, “Monitoring Karenia brevis blooms in the Gulf of Mexico using satellite ocean color imagery and other data,” *Harmful Algae*, vol. 2, no. 2, 2003, pp. 147-160.

[7] M. Tomlinson, R. Stumpf, V. Ransibrahmanakul, E. Truby, G. Kirkpatrick, B. Pederson, G. Vargo, and C. Heil, “Evaluation of the use of SeaWiFS imagery for detecting Karenia brevis harmful algal blooms in the eastern Gulf of Mexico,” *Remote Sensing of Environment*, vol. 91, no. 3,

2004, pp. 293-303.

[8] Y. Suh, L. Jang, N. Lee, and J. Ishizaka, "Feasibility of Red Tide Detection Around Korean Waters Using Satellite Remote Sensing," *J. of Fisheries Science and Technology*, vol. 7, no. 3, 2004, pp. 148-162.

[9] J. Ishizaka, Y. Kitaura, Y. Touke, H. Sasaki, A. Tanaka, H. Murakami, T. Suzuki, K. Matsuoka, and H. Nakata, "Satellite Detection of Red Tide in Ariake Sound, 1998-2001," *J. of Oceanography*, vol. 62, no. 1, 2006, pp. 37-45.

[10] Y. Son, Y. Kang, and J. Ryu, "Monitoring Red Tide in South Sea of Korea(SSK) Using the Geostationary Ocean Color Imager(GOCI)," *Korean J. of Remote Sensing*, vol. 26, no. 5, 2012, pp. 531-548.

[11] Y. Ahn and P. Shanmugam, "Detecting the red tide algal bloom from satellite ocean color observations in optically complex Northeast-Asia Coastal waters," *Remote Sensing of Environment*, vol. 103, no. 4, 2006, pp. 419-437.

[12] Y. Son, J. Ishizaka, J. Jeong, H. Kim, and T. Lee, "Cochlodinium polykrikoides red tide detection in the South Sea of Korea using spectral classification of MODIS data," *Ocean Science J.*, vol. 46, no. 4, 2011, pp. 239-263.

[13] S. Bak, H. Kim, D. H. Hwang, H. Yoon, and W. Seo, "Detection technique of Red tide Using GOCI Level 2 Data," *Korean J. Remote Sensing*, vol. 32, no. 6, 2016, pp. 673-679.

[14] S. Bak, H. Kim, D. Hwang, S. Oh, and H. Yoon, "Red Tide Detection Technique by Using Multi-temporal GOCI Level 2 Data," *Int. J. of Grid and Distributed Computing*, vol. 10, no. 10, 2017, pp. 45-56.

[15] S. Bak and H. Yoon, "Analysis on optical property in the South Sea of Korea by using Satellite Image : Study of Case on red tide occurrence in August 2013," *J. of the Korean institute of Electronic Communication Sciences*, vol. 11, no. 7, 2016, pp. 723-728.

[16] H. Kim, S. Jang, and H. Yoon, "Utilization of Unmanned Aerial Vehicle(UAV) Image for Detection of Algal Bloom in Nakdong River," *J. of the Korean institute of Electronic Communication Sciences*, vol. 12, no. 3, 2017, pp. 457-464.

[17] B. Chae, W. Kim, Y. Cho, K. Kim, C. Lee, and Y. Choi, "Development of a Logistic Regression Model for Probabilistic Prediction of Debris Flow," *The J. of Engineering Geology*, vol. 14, no. 2, 2014, pp. 211-222.

[18] C. Park, Y. Kim, J. Kim, J. Song, and H. Choi, *Data-mining using R.* Seoul: Kyowoo, 2013.

[19] S. Oh, J. Park, and H. Yoon, "Prediction of Red Tide Occurrence by using Oceanic and Atmospheric Data by Satellite," *J. of the Korean institute of Electronic Communication Sciences*, vol. 10, no. 2, 2015, pp. 311-318.

[20] S. Bak, H. Kim, B. Kim, D. Hwang, U. Enkhjargal, and H. Yoon, "Study on Detection Technique for Cochlodinium polykrikoides Red tide using Logistic Regression Model and Decision Tree Model," *J. of the Korean institute of Electronic Communication Sciences*, vol. 13, no. 4, 2018, pp. 777-786.

[21] S. Bak, D. Hwang, H. Kim, B. Kim, U. Enkgjargal, S. Oh, and H. Yoon, "A Study on Red Tide Detection Technique by using Multi-Layer Perceptron," *Int. J. of Grid and Distributed Computing*, vol. 11, no. 9, 2018, pp. 93-102.

[22] P. Kang, H. Lee and S. Cho, "SVM Ensemble Techniques for Class Imbalance Problem," *Proc. of Korea Information Science Society Conference*, vol. 31, no. 2, Korea, 2004, pp. 706-708.

저자 소개

박수호(Su-Ho Bak)

2013년 부경대학교 공간정보시스템공학과 졸업(공학사)
2017년 부경대학교 공간정보시스템공학과 졸업(공학석사)
2018년 현재 부경대학교 대학원 지구환경시스템과학부(박사과정)
※ 관심분야 : 해양 원격탐사, GIS

김흥민(Heung-Min Kim)

2013년 부경대학교 공간정보시스템공학과 졸업(공학사)
2017년 부경대학교 공간정보시스템공학과 졸업(공학석사)
2018년 현재 부경대학교 대학원 지구환경시스템과학부 공간정보시스템공학전공(박사과정)
※ 관심분야 : 해양 원격탐사, GIS

황도현(Do-Hyun Hwang)

2011년 부경대학교 공간정보시스템공학과 졸업(공학사)
2013년 부경대학교 대학원 공간정보시스템공학과 졸업(공학석사)
2018년 현재 부경대학교 대학원 지구환경시스템과학부 공간정보시스템공학전공(박사수료)
※ 관심분야 : 해양원격탐사, GIS

김범규(Bum-Kyu Kim)

2015년 부경대학교 공간정보시스템공학과 졸업(공학사)
2018년 현재 부경대학교 대학원 지구환경시스템과학부 공간정보시스템공학전공(석사과정)
※ 관심분야 : 해양 원격탐사

엥흐자리갈 운자야 (Enkhjargal Unuzaya)

2014년 몽골 과학기술대학교 정보 및 전기통신기술학과 졸업(공학사)
2018년 현재 부경대학교 대학원 지구환경시스템과학부(석사과정)
※ 관심분야 : 무선 통신, 해양원격탐사, GIS

윤홍주(Hong-Joo Yoon)

1983년 부경대학교 해양공학과 졸업(공학사)
1985년 부경대학교 대학원 해양학과 졸업(공학석사)
1997년 프랑스 그르노블 I 대학교 대학원 위성원격탐사전공 졸업(공학박사)
1999년~2002년 여수대학교 해양공학과 교수
2002년~현재 부경대학교 공간정보시스템공학 교수
2012년~2013년 부경대학교 공간정보연구소 초대 소장
2013년 (사)한국클라우드센터럴파크 이사
2014년 한국전자통신학회 부회장
2015년 공간정보 Big Data 센터장
2015년 행정공간정보화연구소 소장
2016년 (사)한국생태공학회 회장
※ 관심분야 : 해양 원격탐사, GIS