2019년도 한국멀티미디어학회 추계학술발표대회 논문집 제22권 2호

빅데이터 기반 적조 예측 시스템 구현

김준한, 강민석, 이동현, 천지연, 윤홍원*
*신라대학교 컴퓨터소프트웨어공학부
e-mail : hwyun@silla.ac.kr

Implementation of Algal Blooms Prediction System based on Big Data Processing

Junhan Kim, Minseok Kang, Donghyeon Lee, Jiyeon Cheon, Hongwon Yun*
*Department of Computer Software Engineering, Silla University

요 약

국립수산과학원의 한국해양자료센터에서 정선해양관측 데이터와 유해적조 데이터를 수집한다. 수집한 데이터를 통합하고 정제한 뒤에 다중 로지스틱 회귀 모델을 이용하여 분석한다. 해양 관측 데이터의 속성별로 유해 적조에 미치는 영향력을 분석하고 유해 적조 예측 시스템을 구현한다.

1. 서론

유해 적조는 해마다 수억에서 수백억에 이르기까지 어업에 종사하는 어민들에게 막대한 재산 피해를 입히고 있다. 적조와 관련된 해양 환경의 분석이나 적조 생물들의 유해성을 평가하는 연구는 있으나 빅데이터를 활용한 적조의 예측 시스템은 거의 없다[1, 2].

우리는 유해 적조 원인 요소의 영향력을 분석하고 적조를 예측하기 위하여 해양 환경 데이터를 수집하고 분석하기로 한다. 국립수산과학원의 해양자료센터에서 정선 해양관측 데이터와 과거의 유해 적조 데이터를 수집한다[3]. 해양 관측 데이터를 다중 로지스틱 회귀로 분석하고 유해 적조 발생을 예측하는 시스템을 구현한다.

2. 데이터셋 수집

국립수산과학원의 해양자료센터에서 2000년 1월 1일부터 2017년 12월 31일까지 지난 17년간의 유해 적조 데이터를 수집한다. 대상 해역은 동해, 서해, 남해, 동중국해이고 수심 0의 데이터를 수집한다. 수집한 데이터셋에는 수온, 염분, 용존산소, 인산연임, 아질산질소, 질산질소, 규산규소 등 7 개의 속성이 들어 있는데, 비어 있는 값이나 신뢰할 수 없는 값들이 들어 있다.

데이터의 필터링과 형식 변환, 결측 데이터의 처리 등을 거쳐서 정제된 데이터를 얻는다. 데이터의 정제를 위하여 파이썬 프로그래밍 언어를 사용하고 CSV형식으로 저장한다.

3. 데이터 분석 및 시스템 구성

해양 관측 데이터의 분석

유해 적조 데이터셋의 종속변수는 이진형으로써 적조의 발생 유무를 나타내고 7개의 속성이 설명변수로 들어 있으므로 다중 로지스틱 회귀 모델로 분석한다[4, 5]. 통계적 분석 결과는 파이썬 프로그래밍 언어와 팬더스 라이브러리를 활용하여 시각화한다. 분석 결과를 살펴보면 수온, 인산염인은 유해 적조의 발생 가능성을 높이고 염분은 발생 가능성을 낮게 하지만, 용존산소, 아질산질소, 질산질소, 규산규소 등은 유해 적조에 미치는 영향력이 낮게 나타난다.

그림 1은 해양 관측 데이터의 분석에 따른 수온과 염분의 로지스틱 회귀선을 보이고 있다. 수온은 적조 발생을 일으키는 요인이고 염분은 적조 발생 가능성을 낮추는 요인임을 나타내고 있다.

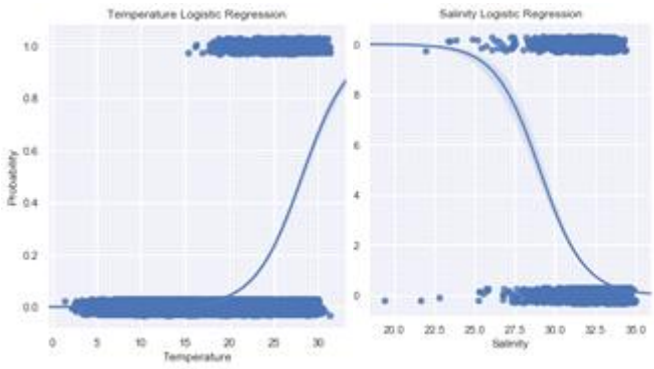

그림 1. 수온과 염분의 로지스틱 회귀선

다중 로지스틱 회귀 분석을 이용하여 속성별 계수를 구하고 다중 로지스틱 회귀식으로 표현하면 다음 식과 같다.

$$f(x) = -5.3 + 0.3x_1 - 0.1x_2 + 0.4x_3$$

위 식에서 -5.3은 절편이고 왼쪽부터 수온, 염분, 질산염인의 회귀 계수이다. 일곱 개의 속성 가운데 나머지 네 개의 속성의 계수는 거의 0에 가깝게 나타나므로 무시한다. 위 식은 유해 적조 발생의 예측에 활용된다.

훈련 데이터와 테스트 데이터에 의한 예측 결과를 보면, 결정한계 50%를 기준으로 했을 때 유해 적조 발생 예측의 정확도는 91.74%로 나타난다.

적조 예측 시스템의 구성

정선 해양 관측 데이터는 정형 데이터로 제공되지만 과거 유해 적조 데이터는 반정형 데이터로 제공되고 있다. 정형 데이터와 반정형 데이터를 통합하고 정제하여 CSV 파일과 데이터베이스 파일을 만든다. 우리가 구현한 적조 예측 시스템의 데이터 분석 모듈은 CSV형식의 해양 관측 데이터셋을 다중 로지스틱 회귀 방법으로 분석하고 시각화한다. 새로운 적조 데이터를 입력하면 적조 발생 확률을 예측하는 본 시스템의 구성도를 그림 2 에 나타낸다.

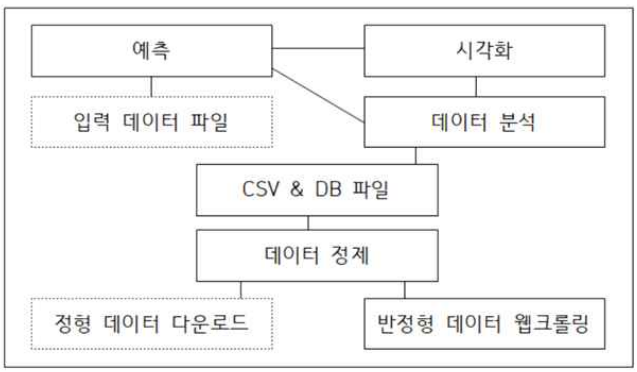

그림 2. 적조 예측 시스템의 구성도

4. 구현

유해 적조 데이터의 분석과 예측을 위하여 다음 세 가지 기능을 PyQt를 이용하여 구현하였다.

- 다중 로지스틱 회귀 분석의 결과 보기
- 선택한 날짜의 해역별 해양 관측 데이터의 평균값 보기
- 적조 발생 확률 예측하기

그림 3. 로지스틱 회귀분석 기본 화면

그림 3은 로지스틱 회귀분석의 기본 화면이며 불러오기를 통하여 CSV 형식의 파일을 가져올 수 있다. 콤보박스를 통해 7가지 속성 가운데 선택할 수 있도록 하였다.

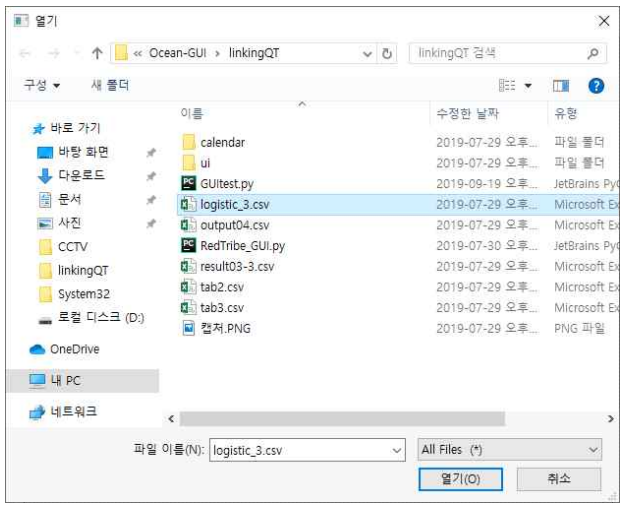

그림 4. 로지스틱 회귀분석 파일 불러오기

그림 4는 분석할 파일을 불러오는 기능이며 QFileDialog의 getOpenFileName()을 사용하였다. QFileDialog는 PyQt5의 위젯이며 간단하게 '불러오기' 기능을 만들 수 있다. 선택한 값은 filename[0]에 저장된다.

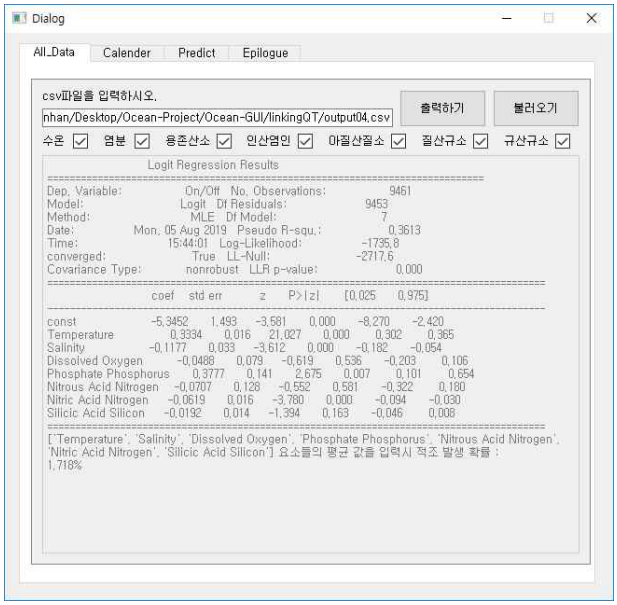

그림 5. 다중 로지스틱 회귀분석 결과 보기

그림 5의 입력 창에서 불러오기 기능을 통해 CSV 파일을 입력하면 다중 로지스틱 회귀 분석을 수행하고 결과를 화면에 나타낸다. 이 기능에서는 분석 대상인 속성을 선택할 수 있고 선택한 속성에 대하여 회귀 분석을 실행한다.

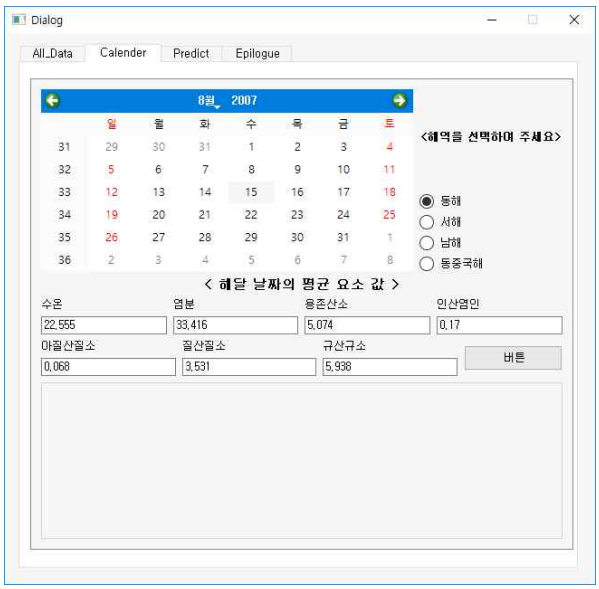

그림 6. 선택한 날짜의 해역별 해양 관측 데이터의 평균값 보기

그림 6은 날짜와 해역을 선택하면 해양 관측 데이터의 속성별 평균값을 보여준다. 라디오 버튼을 이용하여 해역을 중복으로 선택할 수 없도록 하였으며, 캘린더 위젯을 사용하여 달력을 표현하였다.

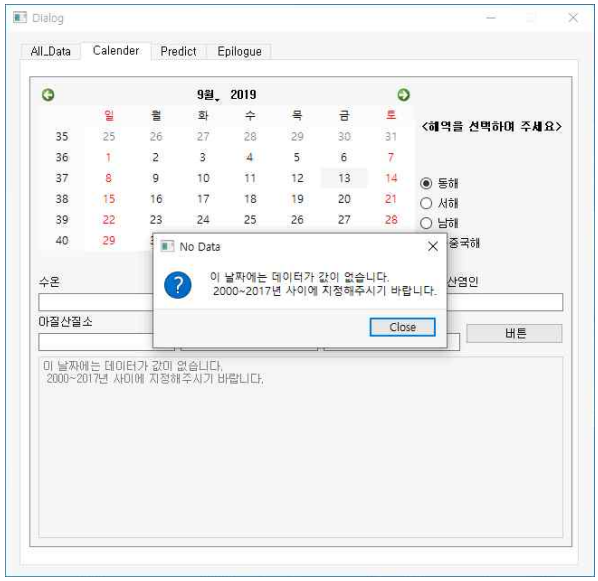

그림 7. 오류 처리 화면

그림 7에서는 해당 날짜와 해당 해역에 데이터가 없을 경우에 오류 메시지를 보여준다. 오류는 exceptKeyError와 IndexError를 통해서 데이터에 값이 없거나 출력이 되지 않을 경우 강제적인 종료를 막기 위하여 사용하였다.

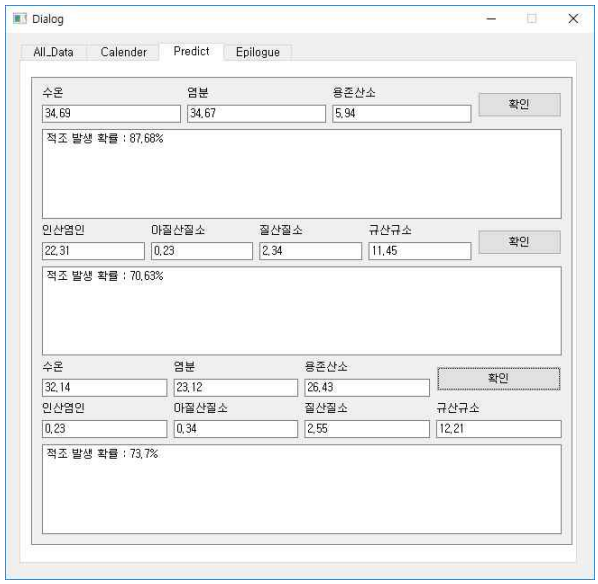

그림 8. 적조 발생 확률 예측하기

해양 관측 데이터를 속성별로 묶어서 예측할 수 있는 기능을 제공하기 위하여 그림 8과 같이 세 개의 창으로 나누었다. 그림 8의 맨 위쪽 창은 수온, 염분, 용존산소 각각에 해당하는 데이터를 입력하면 적조 발생 확률을 보여준다. 가운데 창은 인산염인, 아질산질소, 질산질소, 규산규소에 해당하는 데이터를 입력했을 때 적조 발생 확률을 나타내고, 맨 아래의 창은 모든 속성의 데이터를 입력 했을 때 적조 발생 확률을 나타내고 있다.

statsmodels API를 이용하여 입력 데이터를 받아서 로

직스틱 회귀식(3절의 식 참조)을 이용하여 유해 적조 발생 확률을 예측한다. 위 GUI 프로그램은 윈도우, 맥, 리눅스 등의 모든 운영체제 환경에서 지원된다.

5. 결론

본 연구에서는 유해 적조 발생에 영향을 미치는 요소를 분석하고 적조 발생 예측 시스템을 구현하기 위하여 정선 해양 관측 데이터와 과거 유해적조 데이터를 수집하고 통합하였다. 통합한 데이터를 정제하여 분석 가능한 빅데이터셋을 구성하였다. 해양 관측 데이터셋의 종속변수가 유해 적조의 발생 여부를 나타내는 이진형이고 일곱 개의 독립변수가 있어서 다중 로지스틱 회귀 분석을 이용하였다. 데이터의 수집, 정제, 분석, 구현의 전 과정에 파이썬 프로그래밍 언어와 라이브러리를 활용하였다. 로지스틱 회귀 분석의 결과에 의하면, 수온이 높고 인산염인의 농도가 높을수록 유해 적조의 발생 가능성이 높게 나타났다. 본 연구에서는 적조 예측 시스템을 구현하였다. 적조 예측 시스템은 다중 로지스틱 회귀 분석의 결과 보기, 선택한 날짜의 해역별 해양 관측 데이터의 평균값 보기, 그리고 유해 적조 발생 확률의 예측하기를 제공한다.

참고문헌

[1] 이문옥, 이수화, 김종규, “Cochlodinium polykrikoides 적조발생시의 해양환경적 특징,” 한국해양환경-에너지학회 학술대회논문집, p.222, 2018. 1.

[2]고우진 외, “2017-2018년도 한국연안의 적조발생 상황,” 국립수산과학원, p. 9, 2018.

[3] 국립수산과학원 해양자료센터, http://www.nifs.go.kr/kodc/index.kodc

[4] 위키피디아, https://en.wikipedia.org/wiki/Logistic_regression

[5] DG Kleinbaum, K Dietz, M Gail, M Klein, M Klein, *Logistic regression*, Springer, 2002

[6] M. Chen, S. Mao, and Y. Liu, “Big Data: A Survey,” *Mobile Networks and Application*, Vol. 19, pp. 171-209, 2014.