통계적 기계학습 기법을 이용한 영산강, 섬진강 수계 상수원 지점의 우점 조류 분류에 대한 알고리즘 비교 연구

Comparison of algorithm for dominant algae classification in water source site in Yeongsan, Seomjin river basin Using Statistical Machine Learning technique

황성윤 • 박종환 • 최병웅 • 신동석 / 정강영

환경부 국립환경과학원 영산강물환경연구소 / 환경부 국립환경인재개발원 교육기획과

Introduction

- 영산강 및 섬진강 수계 대표 상수원 지점인 주암호와 탐진호는 지역에 거주하는 시민들에게 물을 공급하는 자원으로서 중요한 역할을 하고 있음. 하지만 `21년 이후 시작된 급격한 강수량 감소와 `22년 극심한 가뭄으로 인해 수자원이 부족할 위기에 처해 있으며 이에 따라 발생하는 우점조류 또한 영향을 받는 것으로 보여짐.
- (연구의 목적) 영산강 수계 상수원 지점의 우점조류 분류를 위한 다양한 통계적 기계학습 알고리즘의 성능을 비교 연구하고 어느 시기에 어떤 조류가 자주 발생하는지 탐색

Data and Methods

- 조사대상지점 : 영산강 및 섬진강 유역 수계 상수원 지점인 주암호(댐앞(J1), 신평교(J2)) 및 탐진호(댐앞(T1), 유치천 합류부(T2))
- 사용변수
  - 수질항목 : BOD, COD, T-N, T-P, TOC, SS, EC, pH, DO, temperature, turbidity(탁도), transparency(투명도), Chlorophyll-a
  - 수리, 수문 : low(저수위), flow1(유입량), flow2(방류량), reservoir(저수량)
  - 반응변수 : dominant(blue(남조류), diatom(규조류), green(녹조류), others(기타조류))
- 분석방법

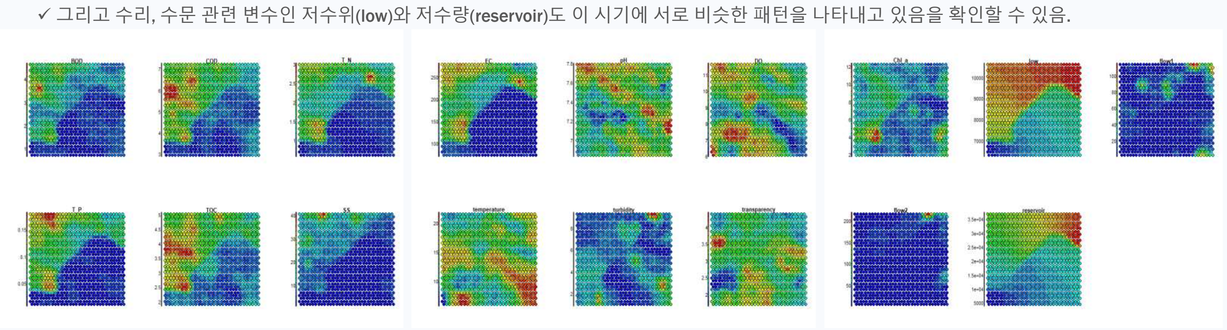
  1) Pattern analysis based Self Organizing Map(SOM)

  2) Compare 11 Statistical Machine Learning algorithm for classification based misclassification rate
- 사용데이터
  - `17년부터 `21년까지의 물환경측정망 주(일)별 자료
  - `17년부터 `20년까지의 자료를 training data로 두고 알고리즘을 훈련시킨 뒤 `21년 자료를 test data로 사용하여 분류 알고리즘에 대한 평가 실시
  - 데이터 출처 : 국립환경과학원 물환경정보시스템 (https://water.nier.go.kr/web)
  - 사용프로그램 : R version 4.2.1

<영산강 수계 지점>

Result and Discussion

- Pattern analysis based Self Organizing Map(SOM)
  - `17~`21년 수질측정망, 조류경보제, 수리, 수문 관련 주(일)별 자료 기반 수질항목 및 수리, 수문 관련 변수에 대한 패턴분석 실시
    - ✓ 수질의 상태를 파악하는 데 필요한 대표적인 수질항목들 중 생물화학적 산소요구량(BOD), 생화학적 산소요구량(COD), 총질소(T-N), 총인(T-P), 총유기탄소(TOC), 부유물질량(SS), 그리고 전기전도도(EC)가 이 시기에 서로 비슷한 패턴을 나타내는 것으로 보아 서로 관련성이 있는 수질항목들끼리는 비슷한 변화를 보인다고 판단할 수 있음.
    - ✓ 그리고 수리, 수문 관련 변수인 저수위(low)와 저수량(reservoir)도 이 시기에 서로 비슷한 패턴을 나타내고 있음을 확인할 수 있음.



- Compare 11 Statistical Machine Learning algorithm for classification based misclassification rate

1) Decision Tree(DT) : 나무모양의 의사결정모형을 통해 특정 변수에 대한 관측값과 예측값을 연결하는 방법. 최적 분리기준 선정을 위한 node의 불순도를 검토 시 회귀의 경우는 오차제곱합($MSE(t) = \frac{1}{n_t}\sum_{i=1}^{n_t}\{y_i(t) - \overline{y_t}\}^2$)을, 분류의 경우는 Gini계수($gini(t) = 1 - \sum_{j=1}^{J} p_j^2(t)$) 또는 Entropy계수($entropy(t) = -\sum_{j=1}^{J} p_j(t) log_2 p_j(t)$)를 사용함.

2) Bagging(Bag) : 다수의 복원추출(sampling with replacement) sample($L_b, b = 1, \dots, B$)을 통한 다수의 decision tree($\vartheta(x, L_b), b = 1, \dots, B$)를 만들고 예측결과를 평균($\vartheta_B(x) = \frac{1}{B}\sum_{b=1}^{B}\vartheta(x, L_b)$)하거나 분류 결과를 바탕으로 다중투표($\vartheta_B(x) = Mode\ \vartheta(x, L_b)$)하여 최종 결론을 도출하는 방법.

3) AdaBoost(Ada) : 예측 성능이 낮은 약한 학습기(weak learner)들을 가중치 조절을 통해 적절히 조합하여 성능이 좋은 강한 학습기(strong learner)를 만드는 방법. 과적합(overfitting)의 위험을 줄이는 장점이 있음.

4) Gradient Boosting(GB) : gradient를 이용하여 모형을 만들고 이를 통해 나오는 잔차(residual)를 다시 모형화하는 과정을 반복하는 방법. 이러한 과정을 통해 편향(bias)을 줄일 수 있지만 과적합(overfitting)의 위험은 높아짐.

5) Random Forest(RF) : Bagging을 실시할 때 각각의 sample에 대하여 임의로 일부의 설명변수를 선택하여 Decision Tree를 만드는 과정을 통해 sample 사이의 연관성을 줄이는 방법.

6) Extreme Gradient Boosting(XGB) : Gradient Boosting 시행 시 나타나는 느린 수행시간과 과적합의 위험을 보완하기 위하여 이 알고리즘에 추가로 병렬학습이 지원되도록 구현한 방법. 자체적으로 교차타당성 검증(cross-validation test)를 수행할 수 있으며 과적합이 나타나는 시점을 감지해주는 Early Stopping 기능이 있음. (본 연구에서는 5-fold cross-validation을 적용하였으며, test data에 대한 mlogloss의 값이 가장 작은 시점을 best iteration으로 판단하였음.)

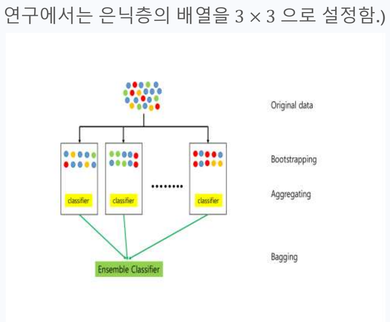
7) Linear Discriminant Analysis(LDA) : R.A.Fisher의 선형판별 경계를 이용하여 분류하는 방법.

8) Flexible Discriminant Analysis(FDA) : spline 방법을 이용하여 비선형(non-linear) 판별 경계를 만들어 분류하는 방법.

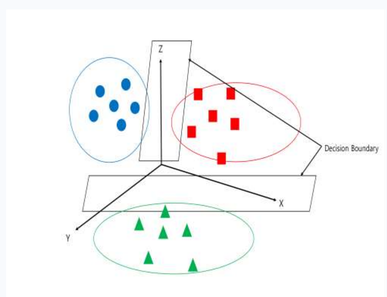
9) Regularized Discriminant Analysis(RDA) : 설명변수(explanatory variable)가 많을 경우 shrinkage와 같은 regularization을 통해 공분산행렬(covariance matrix)에 대한 추정을 향상시켜서 판별 경계를 만드는 방법.

10) Support Vector Machine(SVM) : kernel trick을 통해 decision boundary와 support vector 사이의 거리인 마진(margin)을 최대화하여 분류하는 방법. (본 연구에서는 가장 flexible하다고 알려져 있는 radial basis kernel($k(u, v) =< \varphi(u), \varphi(v) >= \exp[-\gamma||u - v||^2]$)을 적용함.)

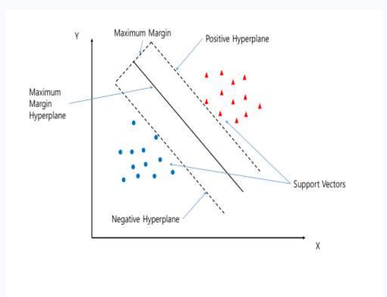
11) Deep Neural Network(DNN) : 입력층(input layer)과 출력층(output layer) 사이에 다수의 은닉층(hidden layer)를 구축하여 만든 신경망 모형. (본 연구에서는 은닉층의 배열을 3×3 으로 설정함.)



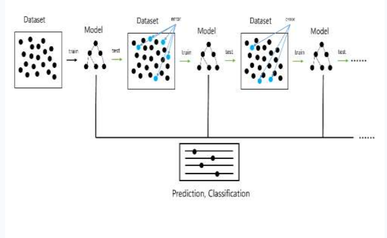
<Bagging>



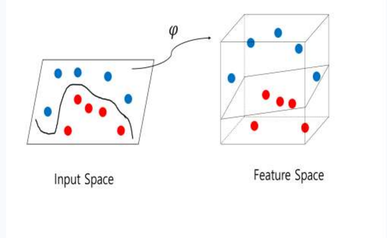
<Discriminant Analysis>



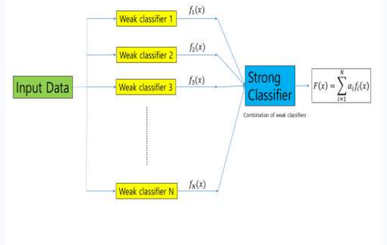
<Support Vector Machine>



<Gradient Boosting>



<Kernel Trick>



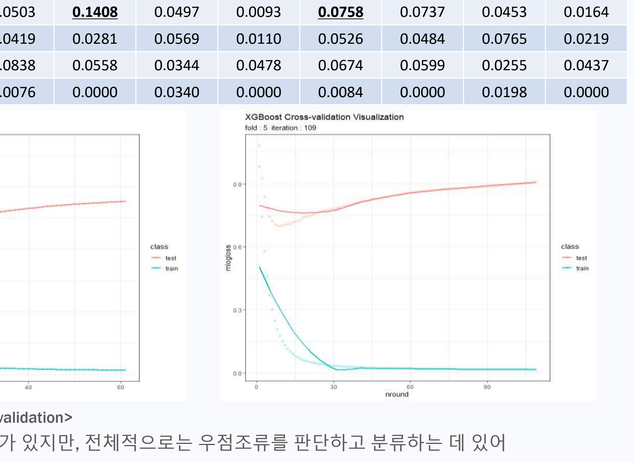
<AdaBoost>

* 나무모형기법(Tree-based technique) 기반 변수 중요도(Variable Importance) 산출 결과

(Gini계수의 감소량이 클수록 변수 중요도는 증가함.)

Algorithm		Bagging				AdaBoost				Gradient Boosting			
site		J1	J2	T1	T2	J1	J2	T1	T2	J1	J2	T1	T2
Explanatory Variable	BOD	1.6355	0.9169	1.1297	**10.8399**	5.6507	2.5200	2.8904	6.3378	4.7210	3.8538	3.8438	7.1205
	COD	3.8333	0.7026	3.8302	0.9498	3.5922	4.2718	5.9266	5.5429	4.0495	2.8880	2.3893	3.2432
	T-N	**7.4599**	3.2770	4.3892	6.6656	7.1632	6.8363	5.1513	**12.6210**	6.4846	4.9263	3.9274	7.1299
	T-P	1.4672	0.9657	1.5616	1.9833	5.6082	5.2904	4.6107	5.4723	3.7721	2.8662	2.9988	3.4984
	TOC	1.2449	1.0364	4.2252	2.1312	2.8110	3.7497	3.3685	5.6915	2.2819	3.0551	2.6374	5.0416
	SS	1.6676	1.8185	2.2088	2.8324	6.9689	5.9100	7.3457	5.8641	4.3569	3.1686	5.0863	**7.8285**
	EC	5.2796	4.0081	**4.9430**	2.3617	5.4458	4.9678	8.5282	6.2782	4.2516	2.9874	5.1073	4.3782
	pH	5.0045	4.8076	0.8526	0.5071	5.3132	**9.8304**	3.6943	5.4650	5.6531	4.7403	1.7300	1.8485
	DO	**27.6322**	4.3681	**38.1803**	0.8516	**10.1666**	8.0513	**9.0549**	3.2185	**18.8630**	**12.5884**	**21.7981**	4.4447
	temperature	**26.8603**	**56.3125**	4.8959	**40.8129**	**8.5444**	**15.2595**	**11.4148**	**7.9387**	**14.1587**	**28.6057**	**9.0861**	**24.0900**
	turbidity	1.3616	**5.6490**	1.9789	0.9841	5.8153	4.4691	7.0753	7.3113	4.0198	5.7638	3.6315	2.8180
	transparency	0.9494	1.0104	0.6039	7.4691	4.8099	3.6236	2.8685	3.0329	2.7471	1.4522	2.0729	3.7890
	Chlorophyll-a	2.7983	3.0667	2.6586	**10.6094**	6.0164	5.7869	6.9407	**11.6879**	4.4275	5.2633	8.6551	**8.4707**
	low	5.9047	**8.5669**	**22.7470**	5.6243	**7.9080**	**8.6515**	6.4177	4.2984	6.1043	**7.7380**	**12.0714**	5.1879
	flow1	2.6420	1.5443	1.2025	1.2009	6.7924	5.7999	**8.5961**	3.8236	5.6886	4.8684	6.4260	3.1708
	flow2	4.0167	1.9053	3.0403	4.1070	6.7450	4.8961	5.1958	5.2944	**7.0044**	3.0835	7.2473	7.4192
	reservoir	0.2423	0.0442	1.5523	0.0696	0.6490	0.0857	0.9204	0.1215	1.4160	2.1512	1.2913	0.5208

Algorithm		Random Forest			
Site		J1	J2	T1	T2
Explanatory Variable	BOD	5.9695	4.6365	3.5421	5.2089
	COD	4.4951	3.5999	3.1710	2.9838
	T-N	6.8724	4.5885	5.3339	**6.8941**
	T-P	5.3997	3.9794	3.3720	3.0757
	TOC	4.0196	4.0364	3.2089	3.7005
	SS	6.3774	4.1817	4.1619	4.2580
	EC	6.3620	4.6390	4.8253	3.2457
	pH	5.2232	5.8216	2.6926	3.0130
	DO	**14.9574**	**13.4656**	**9.9765**	5.0061
	Temperature	**13.7624**	**19.4183**	6.3579	**9.8199**
	Turbidity	5.2893	6.4522	3.4035	3.5417
	Transparency	3.8750	3.6271	2.2659	2.4647
	Chlorophyll-a	5.5955	5.0021	4.4082	**5.3775**
	low	6.5629	**7.0060**	**6.4680**	4.3383
	flow1	5.4151	4.4692	4.0369	2.2046
	flow2	**7.6780**	5.7702	5.4426	4.2021
	reservoir	6.6172	6.5845	**6.4504**	4.2510



<Graph of error in Random Forest>

(Extreme Gradient Boosting의 경우는 3가지 측정 기준(Gain, Cover, Frequency) 기반의 변수 중요도를 산출해준다.)

Extreme Gradient Boosting													
Method		Gain				Cover				Frequency			
site		J1	J2	T1	T2	J1	J2	T1	T2	J1	J2	T1	T2
Explanatory Variable	BOD	0.0449	0.0196	0.0144	0.0817	0.0440	0.0415	0.0052	0.0332	0.0568	0.0530	0.0312	0.0601
	COD	0.0465	0.0173	0.0366	0.0549	0.0345	0.0101	0.0649	**0.1455**	0.0589	0.0276	0.0567	**0.1148**
	T-N	0.0630	0.0491	0.0518	**0.0951**	0.0568	0.0298	0.0496	**0.1691**	0.0589	0.0668	**0.0850**	**0.1257**
	T-P	0.0293	0.0217	0.0260	0.0380	0.0811	0.0110	0.0259	0.0222	0.0632	0.0323	0.0567	0.0437
	TOC	0.0206	0.0185	0.0679	0.0731	0.0259	0.0163	**0.0975**	0.0384	0.0400	0.0369	0.0822	0.0738
	SS	0.0443	0.0275	0.0197	0.0604	0.0571	0.0254	0.0209	0.1181	0.0505	0.0415	0.0510	0.0902
	EC	0.0559	0.0600	0.0780	0.0411	0.0466	0.0340	**0.1631**	0.0173	0.0653	0.0691	**0.1048**	0.0574
	pH	0.0605	0.0619	0.0333	0.0130	**0.0869**	**0.1219**	0.0191	0.0580	0.0695	0.0853	0.0453	0.0410
	DO	**0.2042**	**0.1016**	**0.2582**	0.0059	**0.1613**	0.1071	**0.2418**	0.0034	**0.0989**	**0.1152**	**0.1218**	0.0164
	temperature	**0.1892**	**0.3681**	**0.0871**	**0.2813**	**0.1386**	**0.1811**	0.0581	**0.2365**	**0.1032**	**0.0899**	0.0595	**0.1175**
	turbidity	0.0400	0.0598	0.0230	0.0304	0.0173	0.0702	0.0308	0.0185	0.0484	0.0691	0.0453	0.0492
	transparency	0.0173	0.0126	0.0100	0.0341	0.0159	0.0484	0.0065	0.0168	0.0295	0.0300	0.0255	0.0301
	Chlorophyll-a	0.0258	0.0580	0.0500	**0.1026**	0.0502	0.0785	0.0417	0.0550	0.0526	**0.1014**	0.0680	0.0984
	low	**0.0655**	**0.0689**	**0.1416**	0.0090	0.0503	**0.1408**	0.0497	0.0093	**0.0758**	0.0737	0.0453	0.0164
	flow1	0.0270	0.0297	0.0588	0.0148	0.0419	0.0281	0.0569	0.0110	0.0526	0.0484	0.0765	0.0219
	flow2	0.0579	0.0256	0.0195	0.0644	0.0838	0.0558	0.0344	0.0478	0.0674	0.0599	0.0255	0.0437
	reservoir	0.0082	0.0000	0.0241	0.0000	0.0076	0.0000	0.0340	0.0000	0.0084	0.0000	0.0198	0.0000

<XGBoost Cross-validation>

-> 조사대상지점과 사용한 algorithm에 따라 변수 중요도의 산출결과는 차이가 있지만, 전체적으로는 우점조류를 판단하고 분류하는 데 있어 수온(temperature)과 용존산소량(DO), 그리고 저수위(low)가 다른 변수들과 비교했을 때 더 유의미한 영향력을 보이고 있음을 확인할 수 있음.

* 조사대상지점 및 Algorithm에 따른 test data에 대한 오분류율(misclassification rate) 산출 결과

spot	Classification Algorithm										
	DT	Bag	Ada	GB	RF	XGB	LDA	FDA	RDA	SVM	DNN
J1	0.3846	0.3654	0.3846	0.3654	0.3846	0.4038	**0.3462**	**0.3462**	0.4038	0.3654	0.4038
J2	0.2692	0.2692	0.2692	0.2885	0.2692	0.2885	**0.2500**	**0.2500**	0.2692	0.2885	0.2692
T1	0.1961	**0.1373**	0.1569	**0.1373**	**0.1373**	0.1569	0.3137	0.3333	**0.1373**	**0.1373**	**0.1373**
T2	0.1373	0.1373	**0.0980**	0.1373	0.1373	0.1176	0.2157	0.2353	0.1373	0.1373	0.1373

-> 조사대상지점에 따른 우점조류 분류에 대한 최적의 알고리즘

- ✓ J1(주암호 댐앞) : Linear Discriminant Analysis, Flexible Discriminant Analysis
- ✓ J2(주암호 신평교) : Linear Discriminant Analysis, Flexible Discriminant Analysis
- ✓ T1(탐진호 댐앞) : Bagging, Gradient Boosting, Random Forest, Regularized Discriminant Analysis, Support Vector Machine, Deep Neural Network
- ✓ T2(탐진호 유치천 합류부) : AdaBoost

→ 대체적으로 주암호 관련 지점에서는 Discriminant Analysis 기반 알고리즘이, 탐진호 관련 지점에서는 Tree-based technique 기반 알고리즘이 우점조류를 분류하는 데 좋은 성능을 보였음.

Conclusion

- 오분류율을 기준으로 판단했을 때 우점조류 분류에 대한 최적의 알고리즘은 조사대상지점에 따라 상이한 차이가 있었음. 이는 각 지점마다 조사변수의 분포와 흐름에 차이가 존재하기 때문인 것으로 판단됨.
- 본 연구에 적용된 수질항목 및 수리, 수문 관련 변수 이외에 더 다양한 측정변수를 적용하여 분석한다면 좀 더 신뢰성 있는 데이터에 대한 분석을 통해 우점조류를 분류하는 데 있어 더 좋은 알고리즘을 제안할 수 있을 것으로 여겨짐. 또한 타 수계(한강, 금강, 낙동강) 내의 상수원 지점에 대해서도 본 연구에서 적용한 기법을 적용한다면 해당 수계의 특성에 부합하는 분석 결과를 도출할 수 있을 것으로 기대됨.

2023 사단법인 한국통계학회 THE KOREAN STATISTICAL SOCIETY

국립환경과학원 National Institute of Environmental Research