뜰개 이동경로 예측을 위한 유전 알고리즘 기반 특징선택

김태훈, 문승현, 김용혁*
광운대학교 소프트웨어학부
kth851@naver.com, uramoon@kw.ac.kr, *yhdfly@kw.ac.kr

Feature Selection Based on a Genetic Algorithm for Predicting Drifter Trajectories

Tae-Hoon Kim, Seung-Hyun Moon, Yong-Hyuk Kim*
Sch. Software, Kwangwoon University

요 약

기계학습을 할 때 특징벡터의 선택은 결과에 많은 영향을 미친다. 본 논문에서는 뜰개의 이동경로 예측을 위해 다양한 기계학습 모델을 다루며 뜰개로 1시간마다 관측한 바람, 해류와 OpenDrift 수치모델이 예측한 위치 정보로부터 시계열적인 특성을 고려하여 30가지의 특징을 만들고 이를 입력으로 사용한다. 이때 어떤 특징이 선택되었을 때 가장 좋은 성능을 나타내는지 유전 알고리즘을 포함한 4가지 방법으로 특징선택을 진행하고 2가지 기계학습 모델에 대해 성능평가를 한다.

Ⅰ. 서론

해양사고 발생 시 부유물 및 오염 물질의 이동경로를 빠르고 정확하게 파악해 적절한 대응책을 마련하는 것이 중요하다[1]. 이를 위해 다양한 기상장비로 바람, 해류의 크기와 방향 등 해양 기상정보를 관측하고 전송하는 뜰개와 바다에서 입자의 이동경로를 모델링하기 위해 개발된 프레임워크인 OpenDrift 수치모델이 사용된다[2]. 본 논문에서는 이렇게 얻은 기상정보와 수치모델이 예측한 위치정보로부터 시계열적인 특성을 고려하여 30가지의 특징을 만들고 이를 기계학습 모델의 입력으로 사용하여 뜰개의 실제 관측 위치를 예측한다. 기계학습을 할 때 일반적으로 특징이 많아질수록 좋은 성능을 기대할 수 있지만 학습에 많은 시간이 소요되고 관련 없는 특징이 많아지면 오히려 성능 하락의 원인이 될 수 있다. 이때 특징선택은 학습 연산량을 줄이고 차원의 저주 문제를 완화하며, 예측 성능을 개선하는데 도움을 줄 수 있다[3]. 본 논문에서는 WEKA의 특징선택 필터를 사용하여 특징선택을 하고 선택된 특징들로 2가지 기계학습 모델에 대해서 성능평가 및 비교를 진행한다.

Ⅱ. 본론

1) 실험 데이터

주어진 데이터는 2021년 3월에서 12월 사이 우리나라 해안에서 12개의 뜰개로 1시간 단위로 측정된 23811개의 관측자료를 담고 있다. 데이터는 바람$_u$, 바람$_v$, 해류$_u$, 해류$_v$, OpenDrift가 예측한 위도와 경도, 실제 뜰개가 관측된 위도와 경도로 구성되어 있다. 여기에 시계열 특성을 고려하여 1시간 전부터 현재까지의 값 변화량, 2시간 전부터 현재까지의 변화량, 3시간 전부터 현재까지의 변화량을 계산해 총 30개의 특징벡터를 만들었다. 각 뜰개는 관측 시작 위치가 상이하며, 1개의 뜰개 데이터를 테스트 데이터로 사용하면 나머지 11개 뜰개의 데이터를 학습 데이터로 사용하는 12겹 교차검증 방식으로 학습을 진행했다.

2) 특징 선택

특징선택의 방법은 크게 필터, 래퍼, 임베디드의 3가지로 구분되는데 본 논문에서는 필터 방법과 래퍼 방법을 사용했다. 필터 방법은 각 특징들의 순위를 매겨 특정 임계치 아래의 특징들을 제거하고 연관성 높은 특징들을 선택하는 방식으로 상관계수를 기준으로 0.15 보다 작은 특징들을 제거하고 나머지 연관성이 높은 특징들을 선택했다[3]. 래퍼 방법은 기계학습 모델을 특징들의 조합을 평가하는 목적함수로 사용하여 가장 높은 예측 성능을 보이는 특징 조합을 선택하는 방식으로 선형회귀 모델을 목적함수로 하여 가장 좋은 성능을 보이는 특징 조합을 선택했다[3]. 이때 모든 특징 조합을 다 볼 수 없기 때문에 아무것도 선택되지 않은 상태에서 탐색 기법에 따라 특징을 추가해 나가는 전진 선택 방식으로 특징을 선택했다. 탐색기법은 탐욕 알고리즘 기반 언덕 오르기 탐색기법으로 해를 탐색하는 최상우선탐색(Bestfirst Search)과 유전 알고리즘을 사용해 최적의 해를 탐색하는 유전 탐색(Genetic Search), 평가함수로 모든 특징들을 평가한 뒤 점수가 높은 순서로 특징을 추가하며 해를 탐색하는 순위탐색(Rank Search)이 사용됐다.

각 특징선택 방법에 대하여 기계학습 모델이 예측하는 대상에 따라 4가지 실험이 진행됐다. 뜰개 이동경로 예측을 위한 기계학습 모델은 예측모델과 보정모델로 나뉘는데 예측모델은 뜰개의 현재 위치와 1시간 뒤 위치의 변화량을 직접 예측하고, 보정모델은 OpenDrift 수치모델이 계산한 뜰개의 1시간 뒤 위치를 보정하여 현재 위치와 1시간 뒤 위치의 변화량을 예측한다[2]. 이때 뜰개의 위치는 위도와 경도로 표현되며 위도를 예측하는 모델, 경도를 예측하는 모델, 위도를 보정하는 모델, 경도를 보정하는 모델에 대하여 각각 특징선택이 진행됐다.

3) 성능 검증

특징선택에 대한 성능 검증을 위한 기계학습 모델로 선형회귀와 서포트 벡터회귀 라이브러리(LIBSVM)를 사용했다. 12개의 뜰개 데이터에 대해 각각 선택된 특징들을 입력으로 하는 모델을 만들고 RMSE(Root Mean Square Error)의 평균을 비교해 성능 검증을 진행했다.

4) 실험 결과

[표2]는 4가지 위도와 경도 예측 및 보정 실험에 대해 각 특징선택 방법에 따라 선택된 특징의 평균 개수를 나타낸 결과이다. 필터 방법은 평균 11.4개, 최상우선탐색은 13.3개, 유전탐색은 20.5개, 순위탐색은 23개의

[표1] 유전탐색 특징선택 결과

	1	2	3	4	5	6	**7**	8	9	**10**	11	12	13	**14**	15	16	17	18	19	**20**	21	22	23	24	25	**26**	27	**28**	29	30
위도 예측	**100**	41.7	41.7	83.3	0	41.7	**100**	66.7	100	**100**	50	58.3	58.3	**100**	83.3	41.7	50	8.3	33.3	**100**	75	0	25	16.7	66.7	**100**	75	**100**	66.7	91.7
경도 예측	**100**	100	25	25	16.7	16.7	**100**	75	25	**100**	91.7	83.3	100	**100**	75	33.3	8.3	25	100	**91.7**	0	83.3	41.7	33.3	91.7	**100**	100	**91.7**	100	50
위도 보정	**100**	25	16.7	91.7	8.3	100	**100**	50	100	**100**	25	16.7	100	**100**	91.7	33.3	25	50	58.3	**100**	91.7	8.3	16.7	8.3	100	**66.7**	75	**100**	16.7	100
경도 보정	**91.7**	100	50	16.7	100	25	**100**	58.3	0	**100**	100	75	100	**100**	75	58.3	16.7	25	100	**100**	0	100	25	0	100	**100**	100	**100**	100	41.7

*12겹 교차검증에서 특징별 선택 횟수(단위: 백분율)

[표2] 특징선택 방법별 선택된 특징의 평균 개수

	필터 방법	최상우선탐색	유전탐색	순위탐색
위도 예측	11.83	12.25	19.75	20.75
경도 예측	12.17	14	20.83	23.58
위도 보정	9.25	12.25	19.75	22.92
경도 보정	12.33	14.67	21.58	24.67

특징이 선택되었다. 특징선택에 걸린 평균 시간은 Intel Core i7-7700(3.60GHz), 32GB RAM 기준 최상우선탐색은 113.25초, 유전탐색은 998.75초, 순위탐색은 251.5초가 소요되었으며 성능검증에 걸린 시간은 특징선택을 하기 전 평균 101.48초에서 필터 방법은 58.84초, 최상우선탐색은 51.58초, 유전탐색은 90.48초, 순위탐색은 93.76초가 소요됐다. 4가지 방법 모두 특징 선택을 하기 전보다 검증 과정에서 시간이 적게 소요됐고 선택된 특징 개수에 비례해 시간이 소요됐다.

[표3] 특징선택 성능평가

	기계학습 모델	필터 방법	최상우선탐색	유전탐색	순위탐색	모든 특징
위도 예측	선형회귀	0.003333	**0.001842**	0.001867	**0.001842**	0.003167
	서포트 벡터회귀	0.001842	**0.001833**	0.001850	**0.001833**	**0.001833**
경도 예측	선형회귀	0.004067	0.002150	**0.002117**	0.002125	0.003833
	서포트 벡터회귀	0.002150	0.002133	**0.002092**	0.002108	0.002108
위도 보정	선형회귀	0.005092	**0.001842**	0.001850	0.001858	0.002867
	서포트 벡터회귀	0.004267	**0.001842**	0.001850	0.001858	**0.001842**
경도 보정	선형회귀	0.005575	0.002133	**0.002117**	**0.002117**	0.003833
	서포트 벡터회귀	0.004358	0.002117	**0.002100**	0.002108	0.002108
평균 성능 개선율		-46.90%	20.53%	**20.74%**	20.68%	

[표4] 위도와 경도를 함께 고려한 성능평가

	기계학습 모델	필터 방법	최상우선탐색	유전탐색	순위탐색	모든 특징
위치 예측	선형회귀	0.003700	0.001996	0.001992	**0.001983**	0.003500
	서포트 벡터회귀	0.001996	0.001983	**0.001971**	**0.001971**	**0.001971**
위치 보정	선형회귀	0.005333	0.001987	**0.001983**	0.001987	0.003350
	서포트 벡터회귀	0.004313	0.001979	**0.001975**	0.001983	**0.001975**

[표3]은 4가지 특징선택 방법으로 선택된 특징들로 선형회귀 모델과 서포트벡터회귀 모델을 학습하고 RMSE로 평가한 결과이다. 위도 예측에서는 최상우선탐색과 순위탐색이, 경도 예측에서는 유전탐색이, 위도 보정에서는 최상우선탐색이 가장 좋은 성능을 보였고, 경도 보정에서는 선형회귀 모델에 대해서 유전탐색과 순위탐색이 가장 좋았으며 서포트벡터회귀 모델에 대해서는 유전탐색이 가장 좋았다. 또한, 예측과 보정 모든 실험에 대해 특징 선택을 하기 전과 비교했을 때 상관계수 기반 필터 방법은 평균 -46.90%로 성능이 하락했고, 최상우선탐색은 평균 20.53%, 유전탐색은 평균 20.74%, 순위탐색은 평균 20.68% 성능이 향상됐다.

[표4]는 실제 뜰개 이동경로 예측에서 위도와 경도를 함께 고려해야하기 때문에 [표2]의 결과에서 위도와 경도의 평균값을 계산한 결과이다. 예측에서는 선형회귀 모델에 대해 순위탐색이, 서포트벡터회귀에 대해서는 유전탐색과 순위탐색이 가장 좋은 성능을 보였고, 보정에서는 유전탐색이 선형회귀와 서포트벡터회귀 모두에서 가장 좋은 성능을 보였다. 또한, 이들 모두 성능이 특징선택을 하기 전과 같거나 더 좋아졌으며 평균 21.03% 향상됐다.

[표1]은 4가지 특징선택 방법 중 가장 높은 성능 개선을 보인 유전탐색 기법에 대하여 4가지 위도와 경도 예측 및 보정 실험에서 각 특징들이 몇 번씩 선택되었는지 나타낸 표이다. 단위는 백분율이며 1, 7, 10, 14, 20, 26, 28번 특징인 바람$_v$, 바람$_v$의 1시간 변화량, 해류$_u$의 1시간 변화량, 관측 위도의 1시간 변화량, 수치모델 예측 위도의 2시간 변화량, 해류$_u$의 3시간 변화량, 수치모델 예측 위도의 3시간 변화량은 4가지 실험에서 모두 90% 이상 선택되었다.

Ⅲ. 결론

본 논문은 정확한 뜰개 이동경로 예측을 위해 주어진 데이터로부터 시계열 특성을 고려해 총 30개의 특징벡터를 만든 뒤, 기계학습 모델이 예측하는 대상에 따라 4가지 예측 및 보정 실험에 대하여 상관계수 기반의 필터 방법과 최상우선탐색, 유전탐색, 순위탐색을 통한 래퍼 방법으로 특징선택을 하고 선택된 특징으로 학습한 기계학습 모델에 대하여 성능평가를 진행했다.

실험 결과를 통해 특징선택으로 기계학습에 소요되는 시간을 줄일 수 있음을 확인했다. 또한, 주어진 데이터와 같이 다양한 기상 정보를 다루는 복잡한 회귀문제에서 특징선택을 할 때 상관계수 기반의 필터 방법보다 래퍼 방법이 좀 더 적합함을 확인했고, 평균 성능 개선율을 기준으로 유전탐색 기법을 사용한 래퍼 방법이 특징선택을 하기 전과 비교해 평균 20.74%의 개선율로 가장 좋은 성능을 나타냄을 확인했다.

ACKNOWLEDGMENT

이 논문은 2023년도 해양경찰청 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임(20220463, 지능형 해양사고 대응 플랫폼 구축).

참 고 문 헌

[1] Hyeonki Jeong, Tae-Hoon Kim, Do-Youn Kim, Yong-Hyuk Kim, Seung-Hyun Moon.(2023).Drifter Trajectory Prediction Using Stacked Ensemble with Multiple Machine Learning Algorithms.Journal of Korean Institute of Intelligent Systems,33(5),444-453.

[2] Tae-Hoon Kim, Seung-Hyun Moon, Yong-Hyuk Kim.(2023).Improvement on Predicting Drifter Trajectories Based on Multilayer Perceptron According to Feature Vectors.Korea Artificial Intelligence Conference ,(),116-117.

[3] Girish Chandrashekar, Ferat Sahin. A survey on feature selection methods. Computers & Electrical Engineering, Volume 40, Issue 1, 2014, Pages 16-28.