<결정트리, 앙상블 학습, 랜덤 포레스트 실제 사용 사례 조사 제출>

202184026서동현
201904025김성욱

[결정트리]

<마케팅 분야>

의사결정트리를 사용하여 고객을 유형별로 정의하고 어떤 종류의 제품을 구매할 것인지 예측한다.

<사례>

이병엽, 박용훈, 유재수. (2010). 의사결정트리를 통한 자동차산업의 구매패턴분류. 한국콘텐츠학회논문지, 10(2), 372-380.

의사결정트리 모델을 사용하여 고객 유형별 자동차 구매 패턴을 분류한 결과를 제시하며, 학습된 데이터로부터 자동차 산업의 구매자 행동 및 선호도에 대한 실용적인 제언이나 비즈니스적인 적용 가능성을 논의한다.
본 논문에서는 어떤 고객들이 어떤 차량을 구매하는지, 경소형차와 소형차 구매고객은 어떤 유의한 패턴의 차이를 보이는지의 문제를 정의하였고, 분석의 과정에서 특이하게 도출된 타겟변수들을 재선정하여 세부적인 분류 규칙을 분석하였다.

(1) 어떤 고객이 어떤 차량을 구입하였는가의 분류
결정트리 규칙을 적용하기 위해서 독립변수로는 성별, 연령, 가족형태, 직업, 가족수, 결혼여부, 맞벌이 여부, 신문구독, 월소득이 사용되었고 종속변수로는 문제정의의 분류 타겟인 차종에 대한 데이터로 정의하였다.

표 3. 고객 분류 구매 패턴

Tree Node/Leaf	Target	신뢰도(%)
월소득 (= 260	중형	44.06
월소득) 260	-	-
|직업 = 공무원	대형	39.62
|직업 = 기술숙련직	중형	36.52
|직업 = 자영업	중형	52.15
|직업 = 주부	중형	54.18
|직업 = 무응답	기타	25.00
|직업 = 경영/회사원	-	-
|맞벌이 = O	대형	45.56
|맞벌이 = 무응답	중형	32.30
|맞벌이 = X	-	-
|월소득 (= 450	중형	37.27
|월소득) 450	대형	64.31
직업 = 전문직	-	-
|가족수 (= 3	중형	43.83
|가족수) 3	대형	50.82
직업 = 기타	-	-
|연령 (= 38	중형	42.87
|연령) 38	대형	53.71

[표3] 분석결과를 타켓마케팅 측면에서 분류된 규칙을 정리한 표
: 월소득 260만원 이상인 가정에서 중형차의 구입 가능성이 높음, 맞벌이를 하지 않는 월소득 450만원이상인 고객 분류층에서는 대형차를 선호하는 것으로 분석, 38세 이상의 연령대에서 대형차를 주로 구매하는 패턴이다.

(2) 고객 자동차 구입시 제조사 선택에 있어 어떤 유의한 패턴의 차이를 보이는지 분류하였다.

표 7. "경소형" 자동차를 타겟으로 분류

직업	가족형태	성별	나이	월소득	가족수	신뢰도(%)
기타	미혼/독신	-	-	)135	-	74.03
-	신혼부부	-	-	-	)2	73.10
전문직	미혼/독신	남	)26	-	-	70.01
공무원	미혼/독신	-	)26	-	-	67.35
주부	미혼/독신	-	-	-	-	66.25
-	신혼부부	-	-	(= 125	(=2	64.52
-	자녀성인기	-	)44	-	-	64.52
경영/회사원	가족형성기	-	)44	-	-	63.85
기타	미혼/독신	-	-	-	-	62.20
전문직	미혼/독신	남	)23	-	-	62.05
기타	-	-	-	-	-	60.76

[표7] 경소형 자동차를 구매하는 집단분류의 패턴

가족의 형태로는 미혼/독신, 신혼부부 등의 형태로 경소형 자동차 구매패턴을 볼 수 있고 연령별 분류를 살펴보면 26세 이상, 44세이상에서 경소형 자동차의 구매패턴이 보인다.

표 8. "소형" 자동차를 타겟으로 분류

직업	가족형태	성별	나이	월소득	가족수	신뢰도
전문직	자녀성장기	–	>37 <=40	–	–	88.16
전문직	자녀성장기	–	>37	–	–	73.77
	자녀성장기	–	<=35	–	–	71.85
전문직	가족형성기	–	>30	–	–	70.43
공무원	자녀성장기	–	–	–	–	70.21
전문직	자녀성장기	–	–	–	–	69.62
–	–	–	>48	–	–	68.65
–	자녀성장기	–	–	–	–	67.78
전문직	미혼/독신	–	<=23	–	–	67.49
경영/회사원	미혼/독신	–	–	–	<=2	64.52
전문직	가족형성기	–	–	–	–	62.58
전문직	–	여	–	–	–	62.31
–	–	–	>30	–	–	61.24
기술숙련직	가족형성기	–	–	–	–	61.23

[표8] 소형 자동차를 종속변수로 설정하여 분석

마켓팅 예측 : 소형차, 경소형차의 구매패턴을 살펴보면 전직의 자녀성장기에 있는 그룹에서는 소형자동차의 구매 신뢰도가 상당히 유의하게 도출되었다. 경소형차의 구매 패턴에서는 44세 이상의 집단층에서 경소형차의 구입패턴이 유의하게 도출되었다. 이처럼 의사결정트리를 사용하여 고객의 유형에 따라 자동자 구매 종류를 예측하여 마켓팅에 이용할 수 있다.

● 의학 분야

진찰, 처치, 환자 데이터 등을 변수로 사용하여 혈액 감염을 진단하거나 가슴통증 환자의 심장 발작을 예측함.

● 게임 분야

동작과 얼굴을 인식하는 데 다중의사결정트리 (multiple decision tree)가 사용됨. 사례로 마이크로소프트의 키넥트 플랫폼은 이미지 백만장과 훈련된 트리 세 가지를 사용하여 하루 동안 1000개의 코어로 구성된 클러스터를 사용하여 특정 신체 부위를 분류하였다.

[앙상블 학습]

사례:
강흥식 and 노명규. (2022). 앙상블 학습기법을 활용한 보행자 교통사고 심각도 분류: 대전시 사례를 중심으로. 디지털융복합연구, 20(5), 39-46.

교통사고와 사회·경제적 손실 간의 연계성이 확인됨에 따라 사고 데이터에 기반을 둔 안전 정책 마련 및 중상.사망 등 그 심각도가 높은 교통사고의 절감 방안의 필요성이 제기되고 있다. 본 연구에서는 인구 대비 교통사고 사망자 비율이 높은 대전시를 대상 지역으로 설정하고 보행자 교통사고 데이터를 수집한 후, 기계학습을 통해 최적 알고리즘과 심각도 분류의 주요 인자를 도출하였다. 연구의 결과에 따르면, 적용한 9개 알고리즘 중 앙상블 기반의 학습 기법인 AdaBoost (Adaptive Boosting)와 RF (Random Forest)가 최적의 성능을 보여주었다. 이를 기반으로 도출된 대전시 보행자 교통사고 심각도의 주요 인자는 보행자의 연령이 70대 및 20대이거나 사고유형이 횡단사고에 의한 경우로 나타남에 따라 대전시 보행자 사고 저감 대책을 위한 고려요인으로 제안하였다.

[랜덤 포레스트]

<버퍼거리에 따른 공간 랜덤포레스트를 이용한 월 평균기온 예측 성능비교>

Journal of the Korean Data Analysis Society (October 2020), 22(5), 1809-1818.
https://doi.org/10.37727/jkdas.2020.22.5.1809

김준석, 윤상후

한반도에 발생하는 자연재해의 피해를 줄이기 위해서는 좁고 복잡한 지형적 특징을 고려하여고해상도 기상자료를 생성해야한다. Hengl et al.(2018)이 제안한 공간 랜덤포레스트는 기상정보를고해상도로 공간보간 할 수 있는 기계학습법으로 2D 평면좌표계를 이용한 버퍼거리를 생성하였지만 본 연구에서는 한반도 지형 특성을 잘 반영하기 위해 3가지 버퍼거리를 고려하였다. 고려된 버퍼거리는 평면좌표계(2D80), 구형좌표계(2D84), 그리고 해발고도가 고려된 평면좌표계(3D80)로 계산되었다. 훈련자료는 종관기상관측장비의 2017년 월 평균기온 자료이고 검증자료는 자동기상관측장비에서 수집한 2017년 월 평균기온 자료이다. 예측성능은 평균제곱오차(RMSE), 평균절대오차(MAE)와 결정계수을 기반으로 평가하였다. 예측성능을 평가한 결과 월 효과가 모델에 반영되지 않은 공간 랜덤포레스트가 시공간 랜덤포레스트보다 좋았다. 이는 계절에 따라관측소의 중요도가 상이함을 의미한다. 버퍼거리의 종류에 따른 공간 랜덤포레스트 결과를 살펴보면 버퍼거리로 구형좌표계(2D84)를 이용하고 고도를 입력변수로 사용한 모델이 상대적으로 예측성능이 우수하였다.

-공간 랜덤포레스트-

랜덤포레스트는 의사결정나무의 심화된 기법이며 Breiman(2001)이 제안한 기계학습의 기법중하나로 의사결정나무의 결과를 앙상블 방법(Ensemble method)으로 종합한다. Hengl et al.(2018)은공간적 관계를 설명하기 위해 버퍼거리를 추가한 공간 랜덤분포를 제안하였다. 본 연구에서는공간랜덤포레스트 분석을 위해 R 프로그램(version 3.6.1)의 'GSIF' 패키지를 이용하였다(Hengl et al., 2015).

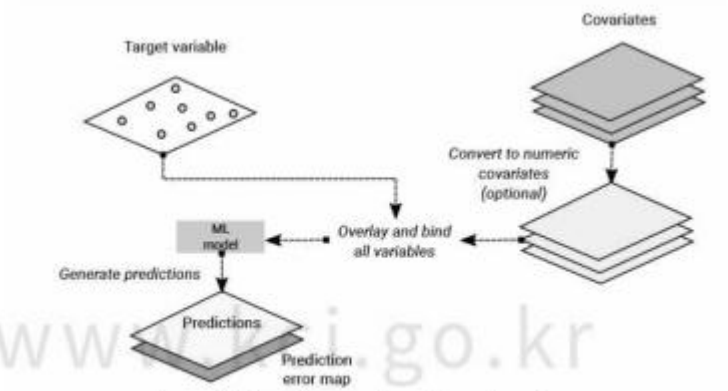


Figure 1. The diagram of spatial random forest

-버퍼거리-

버퍼거리는 격사 사이의 거리로 유클리디언(Euclidean) 방법으로 계산된다.

$$Dist = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2},$$

여기서 갱와 걑는 서로 다른 격자위치를 의미하고 x는 경도, y는 위도, z는 해발고도에 대응된다.
공간 랜덤 포레스트에서는 각 기상관측소의 버퍼거리를 입력변수로 사용하여 공간적 상관성을 훈련시킨다. 따라서 기존의 공간 보간법들과 달리 관측소와의 거리에 따라 가중치가 다르게 부여되므로 예측성능이 좋다. 본 연구에서 사용한 3가지 버퍼거리(2D84, 2D80, 3D80)를 공간 랜덤 포레스트로 훈련 시킨 후 모델의 예측 결과를 비교하였다.

다음의 Figure 2는 버퍼거리를 나타내는 것인데Angle로 표시된 것은 2D84이며 각도를 이용하여 거리를 나타낸다. Projected 2d distance는 2D80이며 높이가 고려되지 않은 평면상의 직선거리이다. Projected 3d distance는 3D80이며 높이가 고려된 직선거리이다.

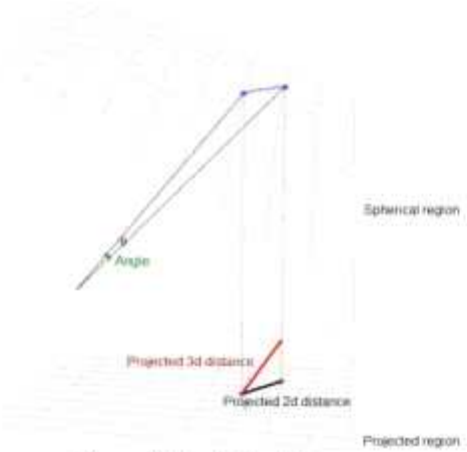
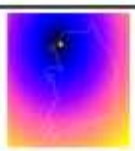
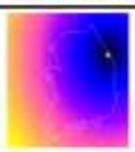
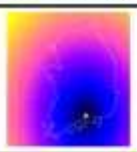
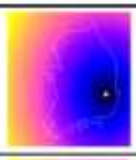
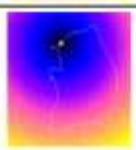
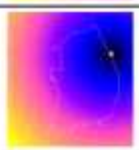
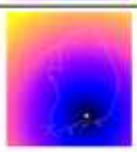
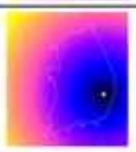
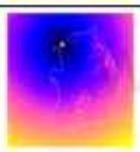
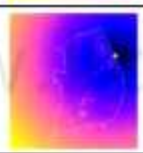
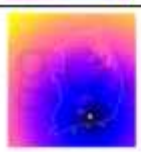


Figure 2. The buffer distances

Table 1. The visualized buffer distances

2D84				...	
2D80				...	
3D80				...	

일부 지점의 버퍼거리를 시각화하면 Table 1이다. 2D84는 위도와 경도의 유클리디언거리이고, 2D80는 미터단위의 평면좌표 거리이므로 단위 규모에 차이가 있다. 2D84는 각도를 이용하여 버퍼거리를 계산하므로 지구의 모양이 고려되어 타원형으로 표현되지만 2D80은 평면거리이므로 원의 형태로 표현된다. 3D80은 고도가 반영되어 산맥의 효과가 버퍼거리에 표현되고 있다.

-모델검증-

모델의 예측성능은 예측값과 실제값의 차이를 계산한 지표인 평균제곱오차(Root Mean SquaredError, RMSE), 평균절대오차(Mean Absolute Error, MAE), 그리고 결정계수(R^2)가 사용하였고 그 식은 아래와 같다.

$$SE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{z_i}-z_i)^2}, \quad MAE = \frac{1}{m}\sum_{i=1}^{m}|\hat{z_i}-z_i|, \quad R^2 = 1-\frac{\sum(\hat{z_i}-\bar{z})^2}{\sum(z_i-\bar{z})^2},$$

여기서 $\hat{Z_i}$는 예측값, Z_i는 관측값, 그리고 $\bar{Z}$는 관측값의 평균이다. RMSE와 MAE는 값이 작을수록 예측성능이 우수하다고 평가하는데 이는 실제 관측값과 예측값의 차이가 작음을 의미하기 때문이다. 예측값과 관측 평균의 차이가 작다면의 R^2값은 커지고 설명력이 높다고 평가한다.

-연구결과-

시공간자료의 랜덤포레스트

시공간자료의 랜덤포레스트를 세우기 위해 3가지 버퍼거리(buffer)와 시간단위인 월(mon)을입력변수로 사용한 모델 A와 고도가 추가로 고려된 모델 B 그리고 위도와 경도가 추가로 고려된모델C를 세웠다. 분석한 결과는 Table 2이다. RMSE와 MAE가 낮을수록 좋은 모델이므로 가장낮은조합을 찾으면 모델 A에 2D84이다. 이는 고도가 높아질수록 온도가 낮아지는 물리적 과정이시공간랜덤포레스트 모델에서는 관측소의 버퍼거리로 대신 설명되기 때문이다. 시공간 랜덤포레스트의버퍼거리 가중치는 계절에 무관하게 반영되므로 월별로 구분하여 공간 랜덤포레스트를 추가로분석하였다.

Table 2. The prediction performance of spatio-temporal randomfores

	buffer + mon.(A)			buffer + mon. + alt.(B)			buffer + mon. + alt. + lon. + lat.(C)		
	2D84	2D80	3D80	2D84	2D80	3D80	2D84	2D80	3D80
RMSE	4.264	4.516	4.453	4.494	4.454	4.437	4.435	4.573	4.546
MAE	3.247	3.408	3.388	3.421	3.388	3.361	3.373	3.488	3.482
R squared	0.849	0.824	0.835	0.831	0.832	0.836	0.835	0.821	0.821

-월별 공간자료의 랜덤포레스트-

모델 A, B, C의 입력변수 중 시간변수인 월(mon)을 제외한 모델 $A^{'}$, $B^{'}$, $C^{'}$로 다시 분석한 결과는 다음의 표이다(Table 3, Table 4, Table 5). 월 단위의 공간자료 분석결과가 시공간자료 분석결과보다 RMSE와 MAE 값이 낮았다. 모델 $B^{'}$버퍼거리 2D84의 결과 중 RMSE는 1~4월, 6~7월, 9월의 평균기온을 다른 모델에 비해 잘 예측하고 있다. 겨울은 지역편차가 커서 RMSE, MAE 결과는 비교적 높게 나타났다. RMSE와 MAE를 기준으로 시공간 랜덤 포레스트와 공간 랜덤 포레스트 모두 버퍼거리 2D84의 예측성능이 좋았다. 이는 지구가 타원형이므로 평면좌표계를 이용하여 버퍼거리를 구하면 가장자리에 가까운 거리는 실제거리보다 길게 추정되기 때문이다.

Table 3. The prediction performance of spatial randomforest(RMSE)

	buffer($A^{'}$)			buffer + alt.($B^{'}$)			buffer + alt. + lon. + lat.($C^{'}$)		
	2D84	2D80	3D80	2D84	2D80	3D80	2D84	2D80	3D80
Jan.	1.677	1.695	1.730	**1.635**	1.680	1.662	1.644	1.646	1.686
Feb.	1.633	1.634	1.634	**1.571**	1.593	1.588	1.587	1.605	1.604
Mar.	1.588	1.597	1.602	**1.506**	1.528	1.518	1.524	1.529	1.517
Apr.	1.497	1.511	1.502	**1.429**	1.439	1.441	1.430	1.437	1.455
May	1.494	1.513	1.518	1.370	1.392	**1.356**	1.379	1.390	1.403
Jun.	1.448	1.465	1.456	**1.374**	1.406	1.411	1.379	**1.374**	1.404
Jul.	1.442	1.458	1.454	**1.376**	1.388	1.388	1.388	1.390	1.398
Aug.	1.401	1.422	1.426	1.366	1.366	**1.345**	1.361	1.376	1.375
Sep.	1.489	1.490	1.496	**1.329**	1.356	1.341	1.352	1.349	1.348
Oct.	1.433	1.450	1.459	1.312	1.351	1.333	1.349	1.362	**1.278**
Nov.	1.623	1.640	1.646	1.605	1.611	1.626	**1.594**	1.600	1.618
Dec.	1.700	1.705	1.696	1.667	1.656	1.683	1.655	1.669	**1.649**

Table 4. The prediction performance of spatial randomforest(MAE)

	buffer(A')			buffer + alt.(B')			buffer + alt. + lon. + lat.(C')		
	2D84	2D80	3D80	2D84	2D80	3D80	2D84	2D80	3D80
Jan.	1.156	1.175	1.191	1.139	1.161	1.151	**1.135**	1.146	1.165
Feb.	1.027	1.028	1.029	0.998	1.007	**0.996**	0.999	1.009	1.006
Mar.	1.025	1.034	1.037	**0.964**	0.975	0.976	0.980	0.983	0.971
Apr.	0.978	0.988	0.979	**0.930**	0.934	0.935	0.935	0.931	0.949
May	1.039	1.055	1.056	**0.961**	0.991	0.964	0.972	0.975	0.988
Jun.	0.966	0.970	0.968	**0.915**	0.936	0.938	0.918	**0.915**	0.938
Jul.	0.920	0.932	0.931	**0.873**	0.884	0.882	0.879	0.886	0.885
Aug.	0.863	0.878	0.883	0.839	0.841	**0.829**	0.841	0.850	0.852
Sep.	1.011	1.020	1.021	**0.913**	0.924	0.928	0.922	0.916	0.916
Oct.	1.010	1.023	1.029	0.923	0.950	0.939	0.951	0.957	**0.909**
Nov.	1.121	1.134	1.139	1.108	1.116	1.125	**1.104**	1.109	1.123
Dec.	1.122	1.135	1.127	1.104	1.101	1.115	1.099	1.111	**1.096**

Table 5. The prediction performance of spatial randomforest(R^2)

	buffer(A')			buffer + alt.(B')			buffer + alt. + lon. + lat.(C')		
	2D84	2D80	3D80	2D84	2D80	3D80	2D84	2D80	3D80
Jan.	0.699	0.694	0.678	**0.716**	0.699	0.704	0.713	0.710	0.697
Feb.	0.647	0.646	0.645	**0.672**	0.665	0.662	0.668	0.657	0.661
Mar.	0.477	0.467	0.470	**0.532**	0.517	0.520	0.523	0.518	0.519
Apr.	0.348	0.341	0.343	0.391	0.390	**0.399**	**0.399**	0.398	0.388
May	0.266	0.249	0.240	0.382	0.387	**0.422**	0.375	0.386	0.379
Jun.	0.446	0.446	0.446	0.515	**0.517**	0.508	0.507	0.514	0.509
Jul.	0.333	0.323	0.316	**0.397**	0.386	0.384	0.385	0.375	0.376
Aug.	0.594	0.572	0.568	0.623	0.610	0.621	**0.625**	0.608	0.604
Sep.	0.586	0.577	0.577	**0.688**	0.663	0.677	0.673	0.672	0.676
Oct.	0.657	0.645	0.648	0.713	0.701	0.707	0.703	0.700	**0.739**
Nov.	0.693	0.676	0.676	**0.699**	0.688	0.682	**0.699**	0.695	0.688
Dec.	0.698	0.693	0.694	0.711	0.711	0.699	**0.715**	0.705	0.713

-결론-

선행연구 Hengl et al.(2018)에서는 공간 랜덤 포레스트의 입력변수로 버퍼거리 2D80만 사용하였다. 한반도는 산지가 70%로 복잡한 지형적 특성을 지니고 있으므로 고도를 반영할 수 있는 버퍼거리(3D80)와 구형좌표계의 버퍼거리(2D84)도 추가로 고려하여 시공간과 공간 랜덤 포레스트 모델을 세웠다. 시공간 랜덤 포레스트로 월평균기온을 보간한 결과 버퍼거리로 2D84를 사용했을 때 결과가 좋았다. 하지만 사계절이 뚜렷한 한반도는 계절에 따라 공간 가중치가 다를 수 있으므로 월단위의 공간 랜덤 포레스트도 추가로 분석하였다. 시공간 랜덤 포레스트에 비해 공간 랜덤 포레스트의 RMSE와 MAE 값이 낮았다. 공간 랜덤 포레스트는 고도와 버퍼거리 2D84가 고려된 모델의 예측성능이 좋았다. 구형좌표계가 평면좌표계보다 우수한 이유는 지구가 타원형이므로 평면좌표계를 이용하면 거리가 멀수록 실제거리에 비해 버퍼거리가 멀게 추정되기 때문이다. 마지막으로 전국521개소에서 구한 버퍼거리 2D84와 고도를 이용한 랜덤 포레스트로 전국의 월별 평균기온을 시각화하여 계절에 따른 한반도 기온특성을 살펴보았다.

버퍼거리에 따른 공간 랜덤 포레스트를 이용한 월 평균기온 예측 성능 비교 1817본 연구에서는 한계점은 다음과 같다. 첫째, 기계학습방법으로 랜덤 포레스트만 고려되었다. 배깅을 이용한 앙상블 말고도 Tak, Jeong, Jung(2019)에서 언급된 경사하강부스팅과 같은 다른 앙상블을고려해 볼 수 있다. 둘째, 입력변수의 선택이 필요하다. 공간 랜덤 포레스트에서 버퍼거리는 관측소의 수만큼 생성되므로 관측소의 수가 증가하면 입력변수가 그만큼 증가한다. 따라서 정확도향상을 위해 버퍼거리의 차원축소 또는 조합선택이 필요하다. 본 연구는 좌표계에 따른 버퍼거리가공간 보간에 미치는 영향을 랜덤 포레스트로 비교분석하였으므로 연구에서 나열한 한계점은 향후연구과제로 남긴다.