지도학습과 비지도학습 비교: 청동기 데이터셋과 역사적 근거를 중심으로

22124688 이윤진

목차

- 데이터셋 소개
- 지도학습
  - 로지스틱 회귀
  - 랜덤 포레스트
- 1차 결론
- 비지도학습
  - K-Means
  - DBSCAN
  - HDBSCAN
- 2차 결론

데이터셋: CODA(마스터 브랜치)

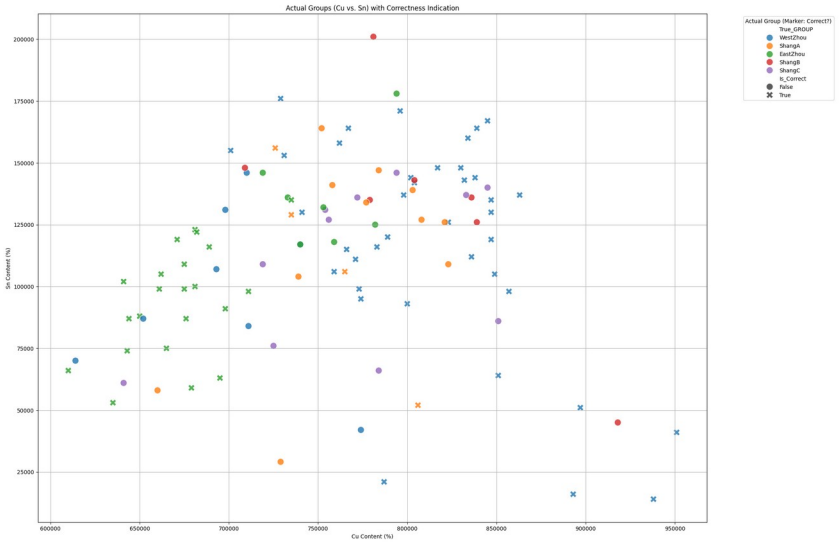
https://github.com/michaelgreenacre/CODAinPractice/tree/master

다양한 주제의 정제된 소규모 데이터

생물학, 고고학 등 다양한 주제 포함

- 일부 데이터는 지나치게 항목이 적음
- 중국 청동기시대 유물의 합금비와 시대/지역에 대한 정보 선택
  - 300여 개의 컬럼
  - 청동 합금비, 왕조와 지역
  - X가 복합적이나 명확한 Y(합금비, 특정 시대의 특정 출토지)

지도학습: 로지스틱 회귀 알고리즘



서주시대의 청동기는 정확도 높음
상나라의 청동기는 정확도 낮음
동주시대의 청동기는 정확도 보통

주나라의 동주 시기 청동기가 발전
- 그러나 제후 중심이기에 파라미터 복잡
- 선형 회귀와 비지도학습에서 성능 나쁨
- 랜덤 포레스트와 비교 요망

상나라는 부족 국가에 근접
- 제대로 된 중심점 부재
- A, B, C 지역 모두 분류 실패

지도학습: 로지스틱 회귀

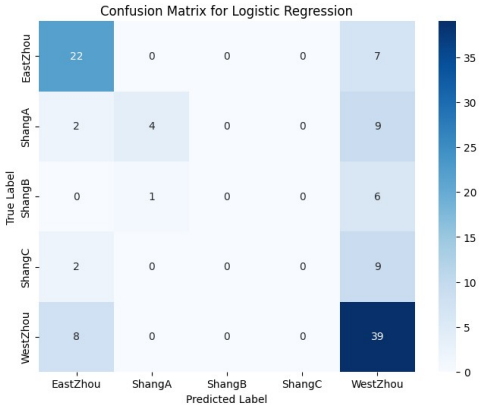
Classification Report:

	precision	recall	f1-score	support
EastZhou	0.65	0.76	0.70	29
ShangA	0.80	0.27	0.40	15
ShangB	0.00	0.00	0.00	7
ShangC	0.00	0.00	0.00	11
WestZhou	0.56	0.83	0.67	47
accuracy			0.60	109
macro avg	0.40	0.37	0.35	109
weighted avg	0.52	0.60	0.53	109

동주의 정밀도 높음
- 비교적 신중하게 분류됨

서주의 정밀도 낮음
- 쉽게 서주로 오판됨
- 상나라 A,B,C 그룹 모두 판독 실패
  - 상나라 A지역은 정밀도는 높으나 분류에 실패
  - 랜덤 포레스트에서 A지역 변화 주목

지도학습: 로지스틱 회귀 혼동행렬



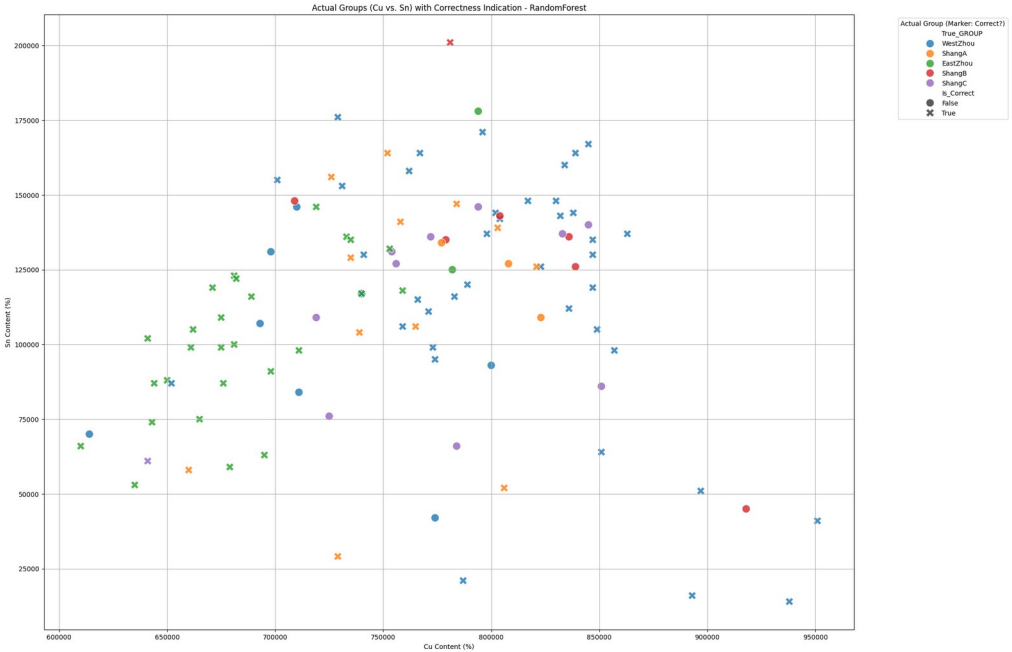
Logistic Regression Coefficients (Feature Weights):

	Cu	Sn	Pb	Zn	Au	Ag	As	Sb
EastZhou	-0.800629	-0.087091	0.329293	0.241459	0.000574	0.366039	-0.192619	-0.214682
ShangA	-0.260075	0.006319	0.504103	0.086957	-0.910918	-0.909930	0.290495	-2.567486
ShangB	0.790993	0.322843	0.842589	-0.012093	-1.247811	0.228282	0.098522	-0.586009
ShangC	0.130292	0.196624	0.223448	-0.099509	0.143418	-0.036140	0.036827	0.543390
WestZhou	0.224472	-0.092215	-1.201968	-0.247543	0.118296	0.187670	0.126386	0.363804

상나라의 분류 정확도는 매우 낮음
서주의 분류 정확도는 높음
- False Positive 다수 보고
- 합금의 특이성이 낮다고 추측 가능
동주 역시 분류 정확도 중간

동주와 서주의 혼동 경향성 있음
상나라 청동기가 서주로 혼동
- 서주는 상나라 바로 다음 시대
  - 문화적 영향력 잔존 가능성

지도학습: 랜덤 포레스트



Shang A:False 비중 대폭 감소
East Zhou: 정확도 대폭 상승

랜덤 포레스트와 같은 다중 분류기는 고고학과 같은 복잡한 데이터에 강점을 보임

지도학습: 랜덤 포레스트

	precision	recall	f1-score	support
EastZhou	0.79	0.93	0.86	29
ShangA	0.80	0.80	0.80	15
ShangB	1.00	0.14	0.25	7
ShangC	0.33	0.09	0.14	11
WestZhou	0.70	0.83	0.76	47
accuracy			0.73	109
macro avg	0.72	0.56	0.56	109
weighted avg	0.72	0.73	0.70	109

상나라 A지역의 정확도 대폭 상승

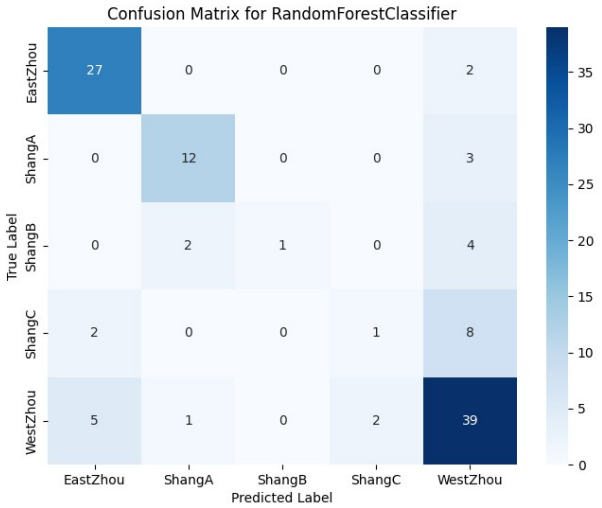
- 예측 불가하던 지역 → 우수한 분류 정확도

전반적인 정확도 개선

서주 그룹의 False Positive 대폭 감소

→ 로지스틱 회귀보다 **복잡한 분류에 적합**

지도학습: 랜덤 포레스트



혼동행렬에서 상나라 A의 성능 개선

- 4→12로 3배 향상

동주 그룹의 성능 개선

22 → 27

서주의 False Positive 감소

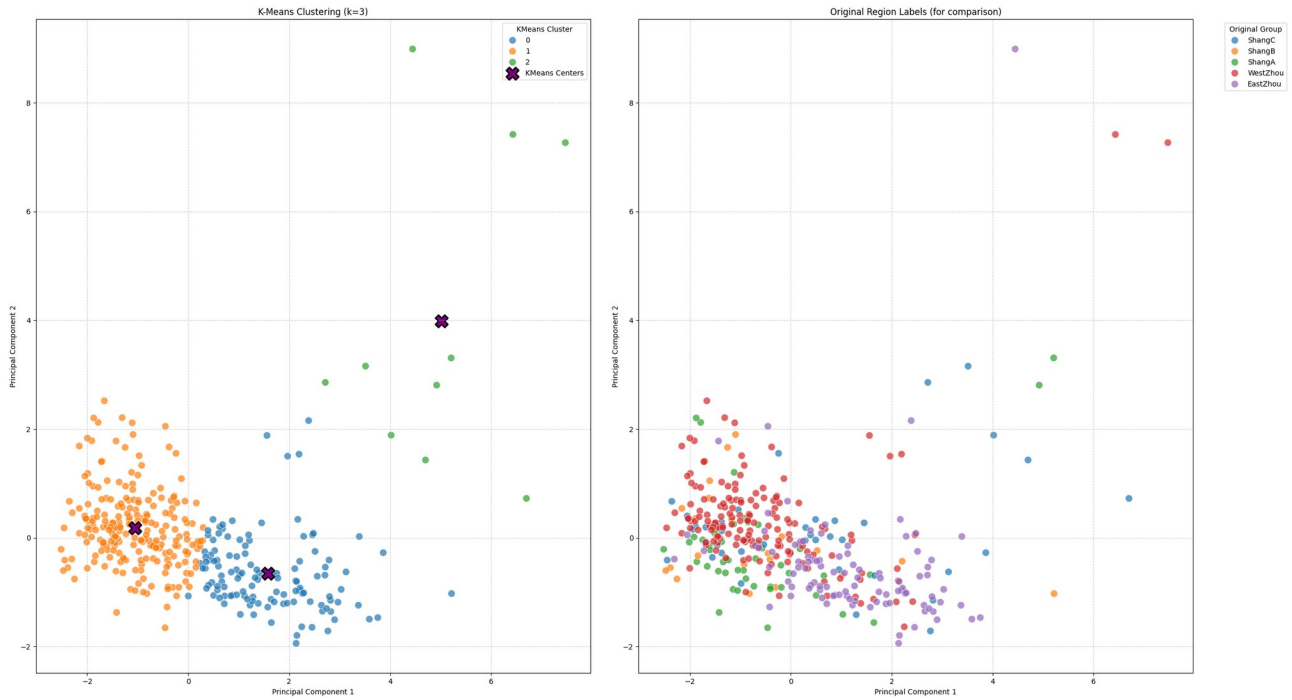
→ 7,9,6,9 → 2,3,4,8

1-6 범위에서 소폭

1차 결론

- 다중 분류기가 단일 분류기보다 복잡한 맥락의 분류에 우수
- 단일 분류기는 복잡한 작업에서 False Positive 등의 혼동이 잦음

비지도학습: K-Means

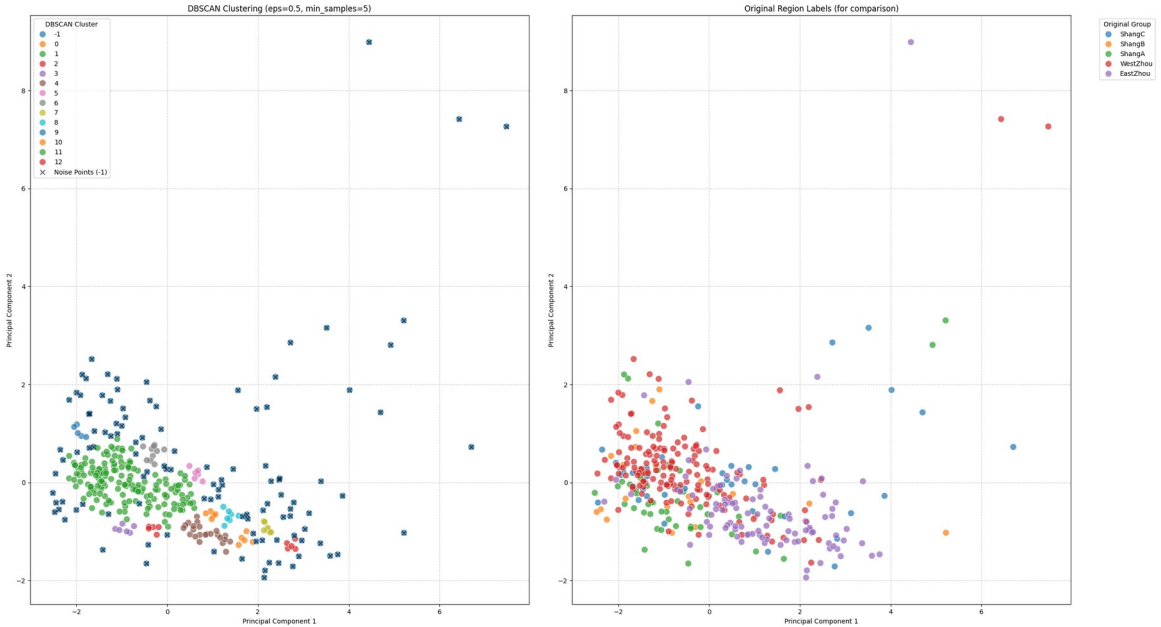


K-Means를 이용한 군집화

- 주성분 분석 기반
  - 주성분으로 X,Y 지정
    - 군집화

K-Means를 쓰기에는
자료의 양상이 복잡

비지도학습: DBSCAN



K-Means보다는 향상

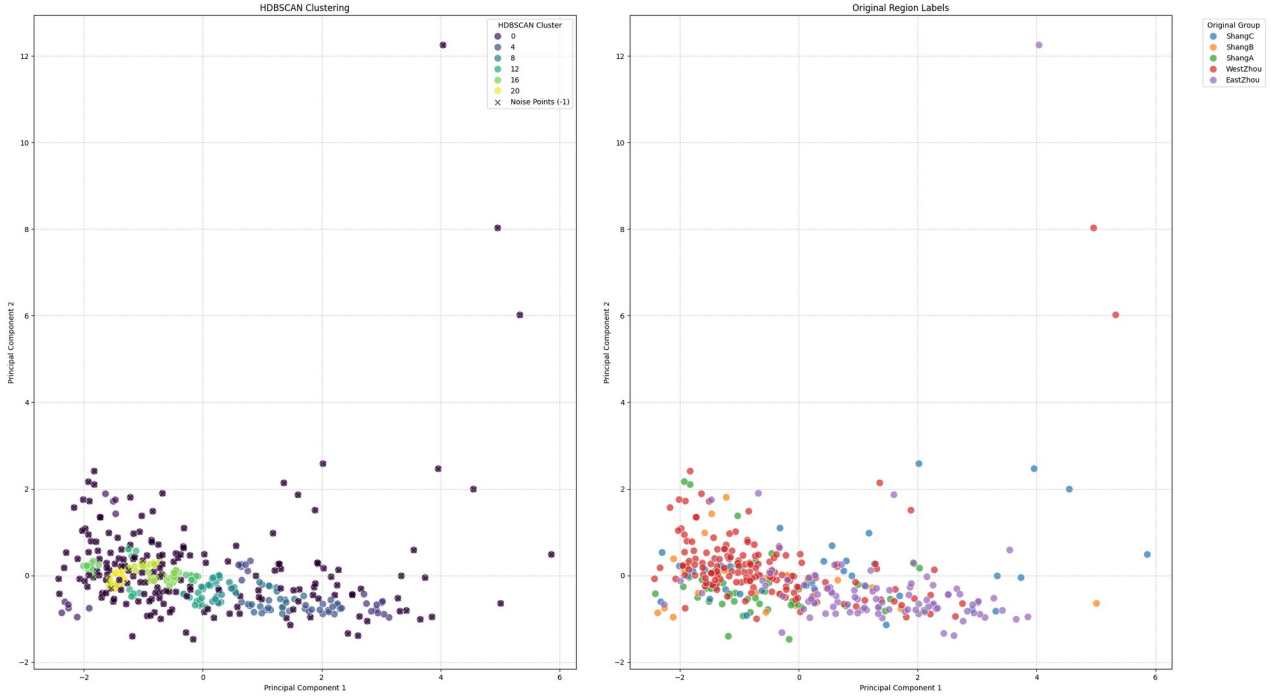
동주시대 분류가 파편화

- 지방별 편차 심한 시대
  - 수동 엡실론이 악영향

여전히 단순 군집화의 경향

- 대략적임

비지도학습: HDBSCAN



비지도학습 중 가장 준수

- 엡실론 수동 지정 X
- Core Point 단독 기반

역시 대략적인 경향성

- 동주, 서주 간의 경계가
  스펙트럼형

최소 클러스터 크기에 민감

2차 결론

- 비지도학습 기반 분류기는 클러스터링이 대략적
- 위치 기반의 분류는 단순히 반경이가까운 그룹을 군집화
- 고고학 데이터와 같은 복잡한 인과가 작용하는 시나리오에서 정상적인 분류 실행 불가