**※ Predictive inference with random forests: A new perspective on classical analyses**

|  |  |
| --- | --- |
| **요약** | ● core model 가정이 위배되었을 때, 발생할 수 있는 문제가 많은데도 거의 대부분은 quantitative 정치학은 inflexible한 회귀 모델(독립변수와 반응변수 간의 선형관계)에 의존  ● nonparametric statistical learning 방법(ex. Random forest)들은 interpretable와 flexible 하다는 장점  ● 최근 연구에 따르면, regularity 조건이 적합할 때 랜덤 포레스트에 의해 하위샘플링된 prediction의 평균은 normal prediction으로 수렴  ● parametric framework의 가설검정과 신뢰구간을 구하는 것처럼 분산 추정을 통해 가설검정과 신뢰구간 구할 수 있음  ● 이 논문은 civil war가 시작된 결정요인에 관한 중요한 연구를 복제함으로써, 분산추정을 통해 가설검정과 신뢰구간을 구할 수 있다는 접근법을 구체화하고  ● 현재 진행중인 연구에 대한 새로운 관점을 통해 nonlinear 관계가 발견될 수 있다는 것을 증명 |
| **주제어** | Random forests, nonparametric, nonlinear |
| **Introduction** | ● flexible statistical & machine learning이 있음에도 불구하고 정치학에서는 고전적인 방법을 통해서만 추론  ● nonparametric ensemble method = Random Forest 는 데이터의 비선형성을 accommodate  ● conventional variance estimate를 통해서 귀무가설의 유의성 검정을 하면서 interpretability의 정도를 설명함  ● 개별 tree(base learner)가 훈련데이터의 하위 샘플로 이루어져 있을 때 R.F 의 통계적 특성들에 대한 연구들 有  - Wager and Athey(2017)은 honesty와 regularity 조건을 따르는 base learner의 특정한 클래스를 주장  \*\* honesty와 regularity 조건: 트리가 마지막 노드를 추정하는데 사용된다는 것을 의미하기 보다는 다양한 데이터에서 트리가 구성된다는 것을 의미  ● Suitably constructed base learners: 점근적으로 정규성에 필요한 조건을 만족하는 tree나 base learner 를 의미함  ● 궁극적으로 R.F 예측의 불확실성을 측정(quantify)할 수 있는 방법을 제공함  ● subsampled random forest에 의한 예측의 분산 추정치는 linear model의 conventional 분산 추정치처럼 변수의 유의성을 검증하는데 사용됨  ● 중요한 것은 결과를 해석하고 추론을 하기 위해 모델의 유연성을 포기해야 한다는 인식은 outdated notion임 🡪 물론, flexible approach는 computational cost ↑   |  |  | | --- | --- | | Random forest | Linear Regression | | - nonparametric method  🡪 particular term-specific parameter의 유의성을 검증 X  🡪 predictions 자체가  conventional parameters를 대신함 | - standard parametric model  - 해석 용이 |   ● 정치학에서 어떤 설명변수가 유의한지를 검증하는 것이 기존의(conventional) 방법은 아니며, 오히려 기존의 연구방법은 모델을 fitting 했을 때 정확도가 높으면 변수가 유의하다고 결론을 내림  ● 현실 데이터에서는 linear model 가정과는 상당히 동떨어져 있음  ● flexible machine learning을 통해 추론할 수 있음  [note] 예측력이 강한 변수가 통계적으로 유의하지 않을 수 있다는 것에 주목 🡪 언제? 독립변수에 대한 그들의 관계가 비선형이고 그 비선형성이 고려되지 않았는데 선형 모델을 적합시켰을 때  ● 방법론 및 data: GDP, fractionalization이 ‘civil war 시작’에 대한 설명력을 강ㅎ |
|  |  |