어떤 의안이 법률이 되는가: 데이터 불균형을 고려한 의안 결과 예측

박은정 epark@dm.snu.ac.kr
강필성 pilsung_kang@korea.ac.kr
조성준 zoon@snu.ac.kr

April 14, 2015

Abstract

법의 제정은 각 개인과 여러 이익집단에 크고 작은 영향을 미친다. 이 때 개개인 또는 이익집단이 법을 수정하거나 제안하는 의안의 결과를 사전에 예측할 수 있다면, 의안의 조속한 통과 내지 폐기를 위해 적극적인 행동을 취하거나, 의안의 집행에 앞서 사전에 대비를 할 수 있다. 본 연구에서는 발의자, 의안 제안 환경 등 의안과 관련한 각종 메타 변수를 도출하여, 대한민국 국회를 대상으로 의안의 결과를 예측해본다. 사용된 알고리즘은 로지스틱 회귀분석(logistic regression), SVM, k-NN, 나이브 베이즈(Naive Bayes), 의사결정나무(decision tree) 등 다섯 가지이다. 법률안이 무조건 통과되지 않는다는 기준 모델에 대해 다섯 개의 알고리즘 모두 정확도가 4% 이상 향상 되었지만, 로지스틱 회귀분석과 SVM 등 특정 알고리즘의 경우 재현율이 낮았다. 한편 데이터의 불균형을 해소해준 후에는 이전 모델에 비해 정확도와 정밀도를 희생함으로써 재현율을 향상시킬 수 있었다.

1 서론

입법은 입법국가에서 중요한 역할을 한다. 법이 어떤 방향으로 제정되느냐에 따라 이익집단에게 끼치는 영향이 달라지기 때문에 많은 단체가 로비를 통해 특정 의안의 입법을 촉구하거나 데모를 통해 반대하는 등 다양한 행동을 취하곤 한다. 입법은 일반적으로 발의자에 의해 의안이 접수된 후, 위원회 심사, 체계자구 심사, 본회의 심의, 정부 이송, 공포의 6 단계를 거쳐 이루어지게 되는데, 각 단계를 거칠 때마다 해당 의안에 대한 논의가 이뤄지고 의안의 대안이 반영되거나 그 자체가 폐기되는 경우도 상당수다. 제 18대 국회의 경우, 결의안, 동의안, 출석요구안 등을 제외하고 철회되지 않은 법률안 중 통과된 의안은 17.4%에 불과하고, 그 중 순수하게 의원이 발의한 법률안만 고려한다면 통과율은 5.91%에 그친다. 뿐만 아니라 각 의안의 통과 여부가 결정되기까지 소요되는 시간은 상당히 길다. 일례로, 18대 국회 의안들이 처리 절차를 거치는데는 평균적으로 470일이었다.

그런데 만약 의안이 발의되는 시점에서 의안 통과의 가능성(likelihood)을 예측할 수 있다면 어떨까? 의안의 통과 가능성이 높은 경우 각 이익 단체별로 조금 더 적극적으로 통과나 철회를 지지할 수 있고, 의안 통과 가능성이 낮은 경우 시간과 비용의 투자를 줄일 수 있다.

한편 정치 전문가가 의안의 결과를 하나하나 예측할 수도 있지만 그림 1에서 보듯 18대 국회에서 발의된 의안은 매해 3,700 건을 초과했고 이는 1대부터 꾸준히 증가가 되어온 수치이다. 현재 19대 국회의 경우에는 의안 발의 건수가 해마다 5,000건을 상회할 전망인데, 이처럼 많은 의안을 전문가가 일일이 감정하는데는 많은 시간과 비용이 소요될 것이다. 만일 감정하는 과정을 자동화할 수 있다면, 정치 전문가는

보다 통과 확률이 높은 의안에 대해 집중적으로 검토할 수 있다. 뿐만 아니라, 데이터를 기반으로 객관적인 예측을 하면 주관성을 배제할 수 있기 때문에, 알고리즘적으로 의안의 결과를 예측하는 것은 정치 전문가의 의견을 보완 내지는 강화하는 효과를 낳을 수 있다.

이런 가능성을 두고 이미 몇몇 연구자가 데이터를 기반으로 의안 처리 과정을 탐색해 보기도 하였다. 의안 원문 기반으로 미국 의회에서 각 의원의 투표 예측을 수행한 [?]는 이상점 모델(ideal point model)에 텍스트를 도입하여 106-111대 의회의 투표 데이터에 대해 89%의 예측 정확도를 얻어 텍스트 기반 예측의 가능성을 확인해주었으며, [4]도 의안의 각종 메타변수와 원문을 활용하여 109-110대 의회에서 의안이 위원회를 통과할지에 대해 평균 90.1%의 예측 정확도를 얻었다. 그러나 이들은 내재적으로 범주 불균형(class imbalance) 문제이다. 실제로 [?]에서는 15%의 표가 yea를 나타냈고 [4]도 12.6%의 의안이 통과하지 못하여, 두 연구 모두 이진 분류에서 일반적으로 사용하는 50% 정확도를 기준 모델로 삼지 않고 85%, 87.4%을 기준으로 모델의 성능을 검증하여 각 4%, 3%의 정확도 향상을 이루었다. 하지만 범주 불균형 문제에서 정밀도나 재현율을 고려하지 않고 정확도만 고려했다는 한계점이 있다.

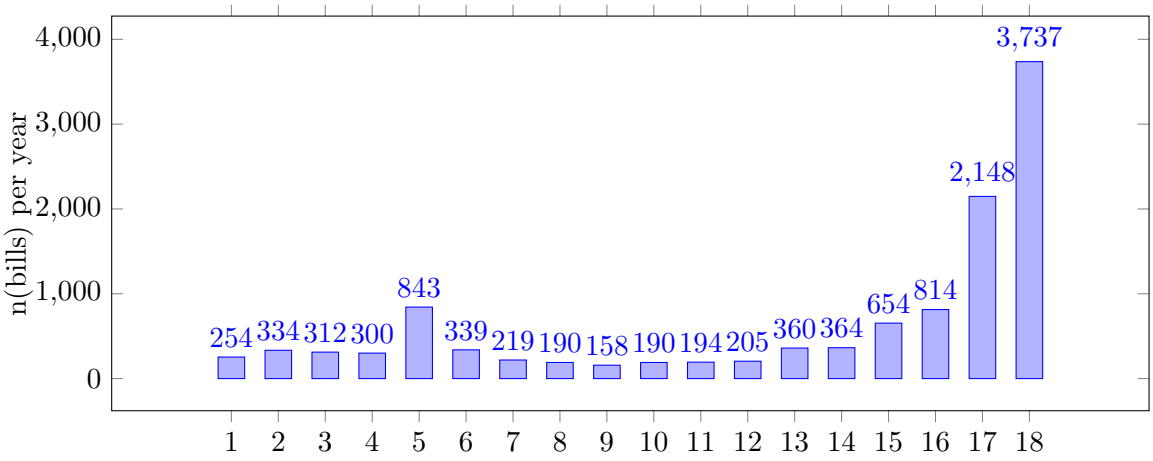

그림 1: 국회 대수 별로 증가하는 의안 발의 개수

본 연구에서는 의안과 관련한 각종 메타(meta) 변수를 도출하여, 대한민국 국회를 대상으로 의안의 결과를 예측해본다. 특히, 의안의 통과 여부는 전형적인 불균형 데이터(imbalanced data)로, 불균형성을 고려한 학습을 하는 것을 주안점으로 삼는다. 구체적으로는 2008년부터 2012년에 걸쳐 18대 국회에서 발의되고 의원들에 의해 철회되지 않은 총 13,405개의 법안이 사용되었으며 사용된 알고리즘은 로지스틱 회귀분석(logistic regression), SVM, k-NN, 나이브 베이즈(Naive Bayes), 의사결정나무(decision tree) 등 다섯 가지 알고리즘을 적용하였다. 발의자, 위원회, 제안 시기, 환경 등에 관한 11가지 종류의 변수를 사용하였는데 대표발의자, 소관위원회 등의 범주형(categorical) 변수를 이진 변수화시키면 총 372개의 변수가 된다. 한편 의안 예측 문제에는 데이터 불균형이 존재하므로, 소수 범주 데이터의 비용을 크게 설정함으로써 데이터 불균형을 해소한 후 로지스틱 회귀분석과 SVM을 이용해 최종적인 예측 결과를 도출하였다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 기존 연구를 살펴본 후, 3장에서 접근 방법을 다룬다. 4장에서 데이터를 설명하고 실험 결과를 다룬 후, 5장에서 결론을 내린다.

2 관련연구

2.1 데이터 기반 의회 분석

데이터를 이용해 의회의 특성을 이해하려는 시도는 국내외에서 다양하게 있었다. 미국 의회의 경우, 의원의 투표 기록을 이용해 의회의 편향성(polarity)를 파악한다든가 [**?**, 3, **?**] 당론이 의원들의 투표에 얼마나 영향을 미치는지 [**?**] 등 투표 기록(roll call)에 대한 다양한 연구가 진행되었다. 특히 [**?**]에서 제안한 D-NOMINATE(Dynamic Nominal Three-step Estimation)은 처음으로 다차원척도법(MDS, multidimensional scaling)을 이용하여 의회 의원들의 이데올로기를 공간상에 전사한 연구로, 정치인 표결 행동에 대한 연구의 기점이 되었다. 이후 이들은 [**?**]에서 DW-NOMINATE(Dynamic, Weighted NOMINATE)를 제안하여 이데올로기의 동적인 변화를 관찰할 수 있는 연구로 확장하였다. 이 연구를 통해 1970년대 이후 미국 양당 간 이데올로기적 거리가 점점 멀어졌다는 것과, 미국 민주당이 "왼쪽"으로 기우는 것보다 공화당이 "오른쪽"으로 기우는 것이 더 심했다는 것을 관찰할 수 있게 되었다. [2]는 미의회 1대부터 110대까지의 투표 데이터에 다층 커뮤니티 탐지(multislice community detection)를 적용하여 시간에 따라 정치인 커뮤니티의 형성이 어떻게 달라지는지 조사하였으며 이를 통해 미의회 15대, 31대, 37대, 74대에 정치인들의 투표 양상에 큰 변화가 있었다는 점을 관찰할 수 있게 되었다.

또, 의안은 보통 한 명의 발의자(sponsor)가 제안하지만, 의안의 표결에 앞서 의안에 강하게 찬성하여 힘을 실어주고 싶은 의원의 경우 공동발의자(cosponsor)가 되기도 한다. 따라서 임의의 두 인물 간 공동발의가 반복적으로 발생할 경우, 두 의원 사이의 이데올로기 유사도가 높다고 할 수 있다. [**?**]는 이 가정을 기반으로 공동 발의 데이터로 네트워크를 구축하였다. 이 연구에서는 공동 발의 빈도와 각 의안의 발의자 수를 이용하여 연결도(connectedness)라는 새로운 지표를 제안함으로써 의원 간 유사도를 계산하였다. 이 지표의 값이 높을수록 입법에서 특정 의원의 영향력이 높다고 할 수 있으며, 결과적으로 이를 통해 어떤 의원의 의안이 더 통과될 확률이 높은지 예측하였다.

미국 의회와 마찬가지로 대한민국 국회도 생성되는 데이터가 상당히 많으며 의안의 원문, 회의록, 의원의 회의 출결 기록, 의안에 대한 본회의 투표 기록 등이 모두 잠재적인 연구 대상이다. [**?**]은 [**?**]와 마찬가지로 공동발의 네트워크를 구축하였다. 이 연구를 통해 한나라당이 야당이었을 당시 열린우리당 소속 의원들보다 강한 영향력을 행사하는 것으로 분석되었다. 그 외에도 법안의 중요도 지표를 제안하고 의원 간 투표 유사도 지표를 만들어 국회를 시각화 하는 등의 시도가 있었다 [**?**, **?**]. 이 연구에서는 18대 국회 본회의 표결에 상정된 총 2,555건의 법안을 분석하여 쟁점법안일수록 한나라당 소속 의원들의 응집력이 민주당 소속 의원들보다 강하다는 점과, 초선 의원들일수록 당론에 부합하는 투표를 많이한다는 점 등을 발견하였다.

이렇게 계산정치학(computational politics) 영역에서는 그동안 설명적(descriptive) 연구가 활발했으며, 그를 토대로 정부와 의회 시스템, 정책 등에 대한 이해가 증진되었다.

2.2 의안 결과 예측

의안의 결과 예측에 관해서는 미의회를 대상으로 하는 연구가 가장 활발하다. 그 중에서 먼저, 이상점 모델(ideal point model) [**?**] 과 베이지안 추론(Bayesian inference) 기반의 생성적(generative) 알고리즘인 IPTM(ideal point topic model)은 의안 원문을 이용해 미의회에서 각 의원의 투표를 예측했다 [**?**]. 이 연구에서는 각 의원 u를 이상점 X_u으로 매핑시키고, 각 의안 d는 난이도(difficulty) A_d와 차별력(discrimination) B_d의 조합으로 나타냈다. X_u, A_d, B_d는 각각 가우시안 선험분포(Gaussian prior)를 적

용하여 추정하며, 특히 A_d, B_d은 의안 원문과 투표 결과를 이용한 sLDA(supervised LDA) 토픽 모델을 응용해서 얻었다. 이 때 투표 결과는 식 1의 랜덤 효과 로지스틱 회귀모형(logistic regression with random effects) $\sigma(t)$을 이용해 분류했으며, 64개의 토픽을 이용한 분류 정확도는 기준 모델 85%에 비해 4% 향상된 89%를 얻었다.

$$p(v_{ud} = 1) = \sigma(x_u b_d + a_d) \tag{1}$$

다음으로, 의안 원문을 비롯하여 각종 메타변수를 이용해 의안이 위원회에서 통과가 될지 예측하는 연구도 있다 [4]. 이 연구에서도 로지스틱 회귀모형을 이용했으며, 메타변수만 이용한 경우 총 3,731개 변수를 이용해 11.8% 오류율, 텍스트 특징까지 활용한 경우 28,411개 변수를 이용해 9.6%의 오류율을 얻어서 기준 모델 대비 약 3%의 정확도 향상을 했다. 텍스트에서는 해당 의안의 카테고리, 위원회의 입장을 추정한 프록시 투표(proxy vote), 의안에서 추출한 BOW(bag of words) 등 세 가지 관점에서 특징을 추출했다.

그 외에도 의원과 의안으로 구성된 이질 그래프(heterogeneous graph) 상에서 랜덤워크(random walks) 모델을 활용하여 과거 투표 데이터 기반의 투표 예측 연구도 있었으며 [**?**], 최근에는 미의회 및 50개 주의 의안의 결과를 예측함과 동시에 원문을 기반으로 의안의 영향을 받을만한 산업 분야를 추정하는 스타트업도 등장했다. 이 스타트업은 자신들의 예측 정확도가 약 93%의 정확도에 이른다고 한다 [**?**].

그러나 기존의 의안 결과 예측 연구들은 의안 예측 문제가 전형적인 데이터 불균형 문제임에도 불구하고, 성능 평가를 정확도(accuracy) 중심으로 했다는 한계를 가지고 있다. 즉 학습모델을 이용해 정확도를 어느정도 향상시켰다 하더라도, 정밀도(precision)나 재현율(recall)은 여전히 낮을 수 있는 것이다. 실제로 [4]에서도 오류율은 낮지만 메타변수만 이용한 경우 F-점수가 0.2343, 텍스트 특징을 활용한 경우 F-점수가 0.4976에 불과했다. 하지만 의안 예측 문제의 경우 소수 범주인 "통과" 여부를 찾는 것이 다수 범주를 찾는 것보다 더욱 중요한 문제이기 때문에 재현율과, 재현율을 고려한 F-지표의 중요성을 간과할 수 없다. 다른 연구들은 정밀도, 재현율, F-점수 없이 정확도만 공개했다.

2.3 데이터 불균형 해소

데이터 기반 의안 예측은 과거 데이터를 기반으로 의안이 통과할 것인지 폐지될 것인지 분류하는 2범주 문제이다. 물론, 폐지되는 다양한 케이스를 고려해서 다범주(multiclass) 문제를 풀어볼수도 있지만 보통은 어떤 방식으로 폐지되는지보다 통과가 되는지의 여부가 주된 관심사이기 때문에 보다 단순한 2범주 문제로 치환하는 경우가 많다. 이렇게 다범주 문제를 2범주로 치환하고 나면, 보통은 통과되는 의안의 수가 폐지되는 수보다 적기 때문에 결정경계(decision boundary)를 보다 정밀하게 탐색하기 위해서는 그 범주간 불균형성을 고려해주는 것이 좋다.

일반적으로 사용되는 방법들로는 언더 샘플링(under sampling), 오버 샘플링(over sampling), 비용 차별(cost-sensitive) 방법 등이 있다. 언더 샘플링은 다수 범주에서 소수 범주의 수만큼 데이터를 샘플링하는 것이고, 반대로 오버 샘플링은 다수 범주의 수에 맞춰 소수 범주의 수를 부트스트래핑(bootstrapping)하는 것이다. 비용 차별 방법은 각 범주의 오분류에 서로 다른 비용을 부과하는 방법으로, 소수 범주에 속한 데이터에 더 큰 비용을 부과하여 오분류되는 것을 최소화한다. 이 외에도 앙상블이나 [**?**] 커널 기반, 액티브 학습(active learning) 기반 방법론들도 다수 제안되었다. [**?**]

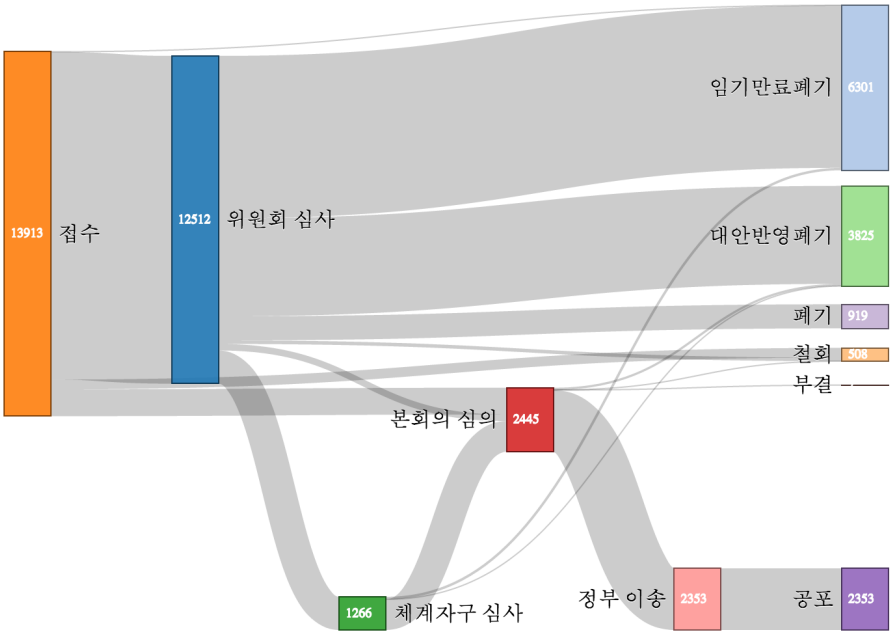

그림 2: 18대 국회 법률안들의 처리 절차를 나타낸 생키 다이어그램. 사각형 모양의 각 노드(node) 안 숫자는 해당 단계를 거친 의안 개수를 의미한다.

3 데이터 불균형을 고려한 의안 결과 예측

의안의 결과를 예측하기 위해서는 먼저 문제의 특성을 잘 이해하는 것이 중요하다. 먼저 의안에는 기존 법률을 수정, 폐지하거나 새로운 법률을 제안하는 법률안만 있는 것이 아니라 결의안, 동의안, 출석요구안, 예산안, 정부 관리직의 임명선출안 등도 포함된다. 그 중에서 일반 시민 및 각종 이익 단체에 영향을 가장 크게 주는 것은 법을 수정하는 법률안이므로, 이 연구에서는 법률안에 중점을 둔다.

다음으로 법률안이 어떤 결과를 가질 수 있는지 파악해보고, "통과"되었다는 것이 무엇을 의미하는지 정의해보자. 그림 2는 18대 국회의 법률안들이 거치는 과정을 도식화한 생키 다이어그램(Sankey diagram)이다.

일반적으로 법률안들은 임기만료폐기, 대안반영폐기, 폐기, 철회, 부결, 공포 등 6가지 상태 중 하나로 끝나게 된다. 그 중에서 공포된 법률안만이 성공적으로 "통과"되었다고 볼 수 있으며, 공포 단계에 이르기 전에 현재 국회의 임기가 만료되거나(임기만료폐기), 같은 목적을 가진 다른 대안이 반영되어 원안이 폐기되거나(대안반영폐기), 발의한 의원들이 철회를 하는 경우 법률안은 법이 되지 못한다. 본 연구에서는 법률안의 발의 시점, 즉 접수 단계에서 해당 법률안이 공포될 것인지의 여부를 예측하는 것을 목표로 한다.

다음으로 데이터의 특성을 보면, 이 문제는 전형적인 데이터 불균형(data imbalance) 문제로, 폐기되는 법률안에 비해 통과되는 법률안이 적다는 특성을 가지고 있다. 제 18대 국회에서 발의된 법률안의 통과율은 17.4%이므로, 법률안은 무조건 폐기된다는 모델을 사용해도 82.6%라는 높은 정확도(accuracy)를 얻게 된다. 따라서 본 연구에서는 정확도 뿐 아니라 정밀도(precision), 재현율(recall), F-점수(F-score)를 계산하여 성능을 점검하며, 법률안이 하나도 통과되지 않는다는 정확도 82.6%, 재현율 0인 모델을 기준 모델(baseline model)로 정한다. 정확도, 정밀도, 재현율, F-점수 각각은 식 (2)-(5)에 나타냈다.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$F - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{5}$$

여기서 TP는 통과된 의안이 통과됐다고 예측한 경우의 수(true positive), FP는 폐기된 의안이 통과됐다고 예측한 경우의 수(false positive), TN는 폐기된 의안이 폐기됐다고 예측한 경우의 수(true negative), FN은 통과된 의안이 폐기됐다고 예측한 경우의 수(false negative)를 나타내며, 이들은 표 1과 같은 혼동 행렬로 나타낼 수 있다.

	Bill pass (predicted)	Bill fail (predicted)
Bill pass (actual)	TP	FN
Bill fail (actual)	FP	TN

표 1: 혼동 행렬

둘째, 발의 데이터에 주기와 변화점(changepoint)이 존재한다. 보통 의안 발의 수는 연말이 다가올수록 발의 의안은 점차 증가하다가 연초에 급락하는 등 주기성을 가진다. 또, 변화점이란 데이터에 영향을 주는 요인이 하나 이상 변화하여 시계열 데이터의 양상이 변하는 것인데, 국회에서도 선거, 의장 변경 등이 의안 발의 및 통과 여부에 영향을 주고 있다. 위의 특성 때문에 의안 발의 시점 또한 의안의 발의와 통과를 이해하는데 무시할 수 없는 요인이 된다.

예측에는 로지스틱 회귀분석(logistic regression), SVM, k-NN, 나이브 베이즈(Naive Bayes), 의사결정나무(decision tree) 등 다섯 가지 알고리즘을 사용한다. 범주형 변수는 1-of-c 코딩을 하여 이진변수화하고, 숫자형 변수는 정규화(normalize)한다.

4 실험

4.1 데이터

본 연구에서는 POODL: POpong Open Data Library[1]에서 제공하는 대한민국 정치 데이터베이스를 이용했다. 이 데이터베이스는 국회 의안정보시스템과 중앙선거관리위원회의 데이터를 수집하여 병합한 것으로, 1대부터 19대까지 국회의원 선거에 출마 또는 당선된 약 13,000명 분에 해당하는 정치인 데이터와 약 54,000건의 의안 정보를 담고 있다 (2014년 기준). 이 연구에서는 POODL에서 제공하는 18대 국회의원, 의안 데이터를 사용했으며, 예측에 사용한 변수는 표 2과 같다.

여기서 사용된 변수들은 [4]에서 제안된 변수를 참고하였으며, 이 연구에서는 특히 대표발의자의 정당, 의안의 소관 위원회, 의안이 발의된 시기 등이 중요한 변수로 꼽혔다. 한편, 위에 나열된 변수 중에는 공동발의자의 수가 있는데 국내에서는 의원 발의 의안의 경우 공동발의자가 최소 10인이 되어야 한다는 점과 입법부에 속한 위원장, 의장, 의원 뿐 아니라 행정부에서 발의할 수 있다는 점이 독특하다.

변수 분류	설명
발의자	대표발의자
	대표발의자의 정당
	대표발의자의 성별
	발의자의 유형 (위원장, 의장, 정부, 의원, 기타)
	공동발의자 수
	발의자 중 다수 정당
	발의자 중 다수 정당의 비율
	발의자 중 제1정당 소속자 비율
위원회	의안의 소관 위원회
기타	발의일의 국회 시작일로부터의 일수
	발의일의 월(month)
	발의일의 국회 연차(1,2,3,4년)

표 2: 고려된 변수들

18대 국회에는 총 14,947개의 의안이 발의되었는데, 그 중에서 13,913개의 의안이 법률 수정안이었고, 13,405개의 의안이 발의자들에 의해 철회되지 않고 남았다. 발의자에 의해 철회되지 않으면서 최종적으로 "통과"가 되어 법률이 된 법률안은 약 17.4%인 총 2,335개였다.

4.2 변수 선택

입법에서 중요한 역할을 하는 변수가 무엇인지 파악하기 위해 단변수 필터 변수 선택법인 (univariate filter variable selection) 카이 제곱 통계량 (chi-square statistics)를 이용해 상위 k개의 변수를 뽑아보았다. 카이 제곱 통계량은 종속변수에 영향력이 큰 독립변수 하나하나가 예측 모형에 기여하는 정도를 계산하는 것으로, 변수간 교호작용은 고려하지 않지만 변수의 중요도를 간단히 도출하기에 좋은 방식이다. 특히, 의안의 메타데이터와 같이 범주형 변수가 많은 경우나, 심지어 데이터가 희소(sparse)한 경우에도 잘 작동하는 것으로 알려져 있어 본 연구의 데이터셋에 적합하게 적용할 수 있다 [**?**]. 카이 제곱 통계량을 도출한 결과 앞서 언급된 변수 중 공동발의자 수가 가장 설명력이 큰 변수로 뽑혔으며 다음으로 발의자의 유형 중 위원장, 의원, 정부 발의 여부, 그리고 발의일의 국회 시작일로부터의 일수, 발의자 중 제1정당 소속자 비율이 순차적으로 등장했다.

4.3 예측 알고리즘

실험은 크게 두 단계로 진행되었는데, 먼저 표 1의 변수들을 이용하여 로지스틱 회귀분석, SVM, k-NN, 나이브 베이즈, 의사결정나무 등 다섯 개의 알고리즘을 이용해 의안의 결과를 예측 해보고, 두번째로 로지스틱 회귀분석과 SVM을 이용해 데이터의 불균형을 처리한 후 결과 예측을 하였다.

4.3.1 로지스틱 회귀분석

로지스틱 회귀분석(logistic regression)은 로짓(logit), 또는 MaxEnt로 불리는 알고리즘으로 종속변수 y가 실수형인 일반 선형회귀분석과는 달리 종속변수가 범주형인 경우를 다룬다. 로지스틱 함수 $\sigma(t) =$

번호	변수
1	공동발의자 수
2	발의자의 유형 (위원장)
3	발의자의 유형 (의원)
4	대표발의자 (친박연대 정영희)
5	발의자의 유형 (정부)
6	발의일의 국회 시작일로부터의 일수
7	발의자 중 제1정당 소속자 비율
8	대표발의자의 정당 (한나라당)
9	의안의 소관 위원회 (규제개혁특별위원회)
10	대표발의자 (통합민주당 최철국)

표 3: 선택된 상위 10개의 변수

$1/1+e^{-t}$를 이용하여 아래의 식 (6)과 같이 x의 범주 y가 1이 될 확률 $F(x)$를 구하며, 파라미터 β_i의 추정에는 최소자승법(least squares)를 사용하는 선형 회귀분석과는 달리 최대우도법(maximum likelihood)을 사용한다.

$$F(x) = p(y = 1|x) = \sigma(\beta_0 + \beta_1 x_1 + ... + \beta_m x_m) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + ... + \beta_m x_m)}} \tag{6}$$

4.3.2 SVM

앞서 거론한 로지스틱 회귀분석은 입력값이 주어졌을 때 출력값에 대한 조건부 확률을 추정하는 생성적(generative) 알고리즘인데 반해, SVM(support vector machine)은 별도의 확률 추정 없이 결과를 직접적으로 추정하는 판별적(discriminative) 알고리즘이다. 기본적으로는 데이터 공간을 고차원의 공간으로 변환하여 범주 간 마진(margin)을 최대화하는 하이퍼플레인을 찾는 2차 최적화(quadratic optimization) 문제를 푼다. 즉, $\mathbf{w}$를 $\mathbf{x}$에 대한 법선 벡터(normal vector)라고 하고 $\mathbf{x}$에 대한 변환 함수 $\Phi(\mathbf{x})$를 정의하면 하이퍼플레인은 $\mathbf{w} \cdot \Phi(\mathbf{x}) - b = 0$로 쓸 수 있다. 또, 비선형 데이터를 수용하기 위해 i번째 개체가 마진을 벗어날 때 페널티 점수를 부여하여 ξ_i라고 하고, 마진과 오분류에 대한 중요도를 정하는 상수 C를 도입하면 최적화 문제에서 목적식은 식 (7)이 되며, 식 (8), (9)의 두 가지 제약조건을 가진다.

$$arg \min_{\mathbf{w},\xi,b} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n} \xi_i \tag{7}$$

$$y_i(\mathbf{w} \cdot \Phi(\mathbf{x_i}) - b) \geq 1 - \xi_i (i = 1, ..., n) \tag{8}$$

$$\xi_i \geq 0 (i = 1, ..., n) \tag{9}$$

여기서 특히 $K(\mathbf{x_i}, \mathbf{x_j}) = \Phi(\mathbf{x_i}) \cdot \Phi(\mathbf{x_j})$를 커널 함수(kernel function)라고 하며, 다항식(polynomial), 시그모이드(sigmoid), RBF 커널 등을 사용할 수 있다. 본 논문에서는 RBF 커널을 사용하였다.

4.3.3 k-NN

거리상으로 가장 가까운 k개 점을 기반으로 투표하는 것으로, 별도의 학습모델을 구축하지 않는다. 즉, 각 범주를 l이라하면 y에 대한 예측값은 다음의 식 (10)과 같이 구할 수 있다.

$$\hat{y}(x) = arg\max_{l} \sum_{x_j \in N(x)} I(y_j = l) \tag{10}$$

4.3.4 나이브 베이즈

나이브 베이즈(naive Bayes)는 베이즈 이론(Bayes Theorem)에 기반한 간단한 확률모형으로, 변수들 간 독립성(independence)를 가정하는 것이 특징적이다. y에 대한 예측값은 다음의 식 (11)를 따른다.

$$\hat{y}(x) = arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i|y) \tag{11}$$

4.3.5 의사결정나무

의사결정나무(decision tree)는 특정 지표를 기준으로 데이터를 가장 잘 가르는 점을 반복적으로(iterative) 찾는 알고리즘이다. 자주 사용되는 지표로는 식 (12)의 지니 불순도(Gini impurity) 또는 식 (13)의 엔트로피(entropy) 등이 있다. 여기서 p_k는 k 번째 범주로 레이블링된 데이터의 비율이다. 본 연구에서는 이 중 지니 불순도를 사용하여 결과를 도출하였다.

$$I_G(A) = 1 - \sum_{k=1}^{m} p_k^2 \tag{12}$$

$$I_E(A) = -\sum_{k=1}^{m} p_k^2 \log_2(p_k) \tag{13}$$

4.4 실험 결과

먼저, 앞의 다섯 가지 알고리즘을 이용하여 의안을 예측한 첫 번째 실험에 대한 결과는 표 4와 같다. 과적합을 방지하기 위해 실험별로 10-fold cross validation을 적용하여 각 지표에 대한 평균값을 기록하였다. 알고리즘의 정확도는 86.8%에서 91.1%로 기준 모델의 82.6%에 비해 전반적으로 4% 이상 상승하였고 의안 예측 연구에서 가장 활발하게 사용되는 로지스틱 회귀분석이 가장 높았지만, 알고리즘 간 편차는 크지 않았다. 한편, 로지스틱 회귀분석과 SVM의 재현율은 40-50% 수준으로, k-NN이나 나이브 베이즈, 의사 결정 나무에 비해 재현율이 20%이상 떨어지는 반면 정밀도는 높게 도출되었다. 결과적으로 알고리즘 간 F-점수는 큰 편차가 없었고, 의사결정나무가 가장 높은 F-점수를 얻었다.

두 번째 실험은 데이터의 불균형을 처리한 후 결과 예측을 하는 방식이다. 불균형은 학습 데이터에서 각 범주(class)의 비율에 따라 소수 범주의 오분류 비용을 가중함으로써 해소했다. 알고리즘은 앞의 다섯 가지 알고리즘 중 정확도가 가장 높은 두 가지 알고리즘이자, 의안 연구에서 가장 빈번하게 활용되는 로지스틱 회귀분석과 SVM을 사용했다. 이 실험에 대한 결과는 표 5과 같다. 알고리즘의 정확도와 정밀도는 앞선 실험에 비해 다소 떨어졌지만, 재현율은 각각 55.2%에서 74.9%와, 43.9%에서 74.93%으로 크게 상승하여 데이터 불균형 보정의 가능성을 확인할 수 있었다. 실제로 의안 예측에서는 통과되지 않는 의안을 통과된다고 하는 것보다 통과되는 의안을 통과되지 않는다고 하는 것의 비용이 크기 때문에 정밀도보다

분류기	정확도(accuracy)	정밀도(precision)	재현율(recall)	F-점수
Logistic regression	0.9106	0.9107	0.5521	0.6601
SVM	0.9018	0.9950	0.4394	0.5596
k-NN	0.8683	0.6047	0.7493	0.6613
Naive Bayes	0.8717	0.6073	0.7438	0.6645
Decision tree	0.8786	0.6281	0.7447	0.6773

표 4: 다섯 가지 알고리즘의 의안 결과 예측 성능

재현율을 높이는 것이 중요하다고 할 수 있다. 한편, 로지스틱 회귀분석의 경우 앞서 F-점수 기준으로 가장 성능이 좋았던 k-NN과 비교해도 약간의 성능 향상을 얻을 수 있었다.

분류기	정확도(accuracy)	정밀도(precision)	재현율(recall)	F-점수
Logistic regression	0.8794	0.6261	0.7487	0.6791
SVM	0.8640	0.5828	0.7493	0.6512

표 5: 데이터 불균형을 해소한 후 두 가지 알고리즘의 의안 결과 예측 성능

4.4.1 소관 위원회별 예측

다음으로 가장 좋은 성능을 나타낸 로지스틱 회귀분석을 중점으로 소관 위원회별 예측 성능을 심층적으로 조사했다. 18대 국회에서 등장한 소관 위원회 37개 중 상임위원회나 특별위원회와 무관하게 소관 법안의 개수가 100건이 넘어서 데이터가 충분하다고 판단되는 위원회 17개에 대한 성능 지표를 표 6에 나타냈다.

총 17개 위원회 중에서 10개의 위원회에서 범주 불균형을 보정하기 이전보다 좋은 F-점수를 얻을 수 있었다. 한편 범주 불균형 보정 전의 F-점수가 나은 경우도 등장했지만, 이들의 재현율을 비교해보면 불균형을 보정하기 전보다 후에 재현율이 상승한 것을 발견할 수 있었다. 의안 통과 예측 문제에서는 소수 범주에 속하는 통과되는 의안을 찾아내는 것이 더 중요하기 때문에 이와 같이 재현율이 높아지는 것이 더 의미있는 결과이다.

4.4.2 국회 연차별 정확도 및 정밀도 비교

마지막으로 로지스틱 회귀분석의 연차별 예측 성능은 표 7과 같았다. 마찬가지로 4년차를 제외하고는 보정 후 F-점수가 더 높았다.

5 결론

본 연구에서는 18대 국회에서 발의된 법률안의 다양한 속성을 이용해 법률안이 발의된 시점에서 다섯 가지 알고리즘을 이용하여 통과 여부를 예측하였다. 법률안이 무조건 통과되지 않는다는 것을 기준 모델로 삼고도 다섯 개의 알고리즘 모두 4% 이상의 정확도 향상을 통해 86% 이상의 정확도를 내었지만, 재현율이 낮았다. 한편 데이터의 불균형을 해소해준 후에는 이전 모델에 비해 재현율과 F-점수를 높일 수 있었다. 하지만 F-점수가 높아지지 않거나 오히려 안 좋아지는 경우도 있었는데, 이 경우에도 의안 예측에서 재현율은 올라간 것으로 확인할 수 있었다. 의안 예측 문제에서 통과된 의안을 놓치지 않는 것을

위원회명	기존 정확도	정밀도	재현율	F-점수	보정 후 정확도	정밀도	재현율	F-점수
정치개혁특별위원회	0.9963	1.0000	0.9412	0.9697	0.9963	1.0000	0.9412	**0.9697**
국회운영위원회	0.9762	1.0000	0.7308	0.8444	0.9830	0.9565	0.8462	**0.8980**
법제사법위원회	0.8723	0.9588	0.4326	0.5962	0.8997	0.7458	0.8186	**0.7805**
보건복지가족위원회	0.8868	0.8800	0.7097	**0.7857**	0.8585	0.7667	0.7419	0.7541
보건복지위원회	0.9815	1.0000	0.7097	**0.8302**	0.9685	0.7765	0.7097	0.7416
외교통상통일위원회	0.8475	0.9000	0.4186	0.5714	0.8757	0.7561	0.7209	**0.7381**
여성가족위원회	0.9271	0.9500	0.5278	0.6786	0.9190	0.7222	0.7222	**0.7222**
기획재정위원회	0.9228	0.8571	0.5455	0.6667	0.9099	0.6441	0.8128	**0.7187**
지식경제위원회	0.8622	0.8206	0.7349	**0.7754**	0.7828	0.6250	0.8233	0.7106
행정안전위원회	0.9373	0.8919	0.6027	**0.7193**	0.8972	0.5776	0.8493	0.6876
국토해양위원회	0.9103	0.9043	0.5882	**0.7128**	0.8592	0.5964	0.7924	0.6805
교육과학기술위원회	0.9150	0.9429	0.4925	0.6471	0.8855	0.6121	0.7537	**0.6756**
국방위원회	0.8484	0.9565	0.3492	0.5116	0.8375	0.6250	0.7143	**0.6667**
환경노동위원회	0.9154	0.9610	0.5175	**0.6727**	0.8566	0.5517	0.7832	0.6474
농림수산식품위원회	0.8360	0.9434	0.4587	0.6173	0.7910	0.6485	0.6009	**0.6238**
정무위원회	0.9270	0.9195	0.5797	**0.7111**	0.8584	0.5313	0.7391	0.6182
문화체육관광방송통신위원회	0.9071	0.9091	0.4800	**0.6283**	0.8626	0.5794	0.5840	0.5817

표 6: 범주별 비용 보정 전후의 소관 위원회별 의안 결과 예측 성능 비교 (보정 후 F-점수 기준 내림차순)

연차	기존 정확도	정밀도	재현율	F-점수	보정 후 정확도	정밀도	재현율	F-점수
1년차	0.7885	0.4586	0.8655	0.5799	0.6336	0.6540	0.8534	**0.6437**
2년차	0.9483	0.4840	0.9128	0.6409	0.6504	0.7533	0.8952	**0.6981**
3년차	0.9651	0.4989	0.9306	0.6577	0.5772	0.7675	0.8938	**0.6589**
4년차	0.9949	0.7984	0.9595	**0.8859**	0.6598	0.8493	0.8840	0.7427

표 7: 범주별 비용 보정 전후의 연차별 의안 결과 예측 성능 비교 (보정 후 F-점수 기준 내림차순)

더 중요한 문제라고 판단한다면, 이는 바람직한 결과라고 할 수 있다. 또한, 기존 연구가 대부분 로지스틱 회귀분석을 적용한데 반해, 다양한 알고리즘을 도입해 알고리즘 간 성능 비교도 할 수 있었다. 뿐만 아니라 미국 의회와는 다른 특성을 가지는 한국 국회에 대해 실험을 진행한데 의의가 있다.

앞으로 이 연구를 발전시켜 의안이 여러 심사 단계 중 어느 단계에서 탈락할 것인지를 예측해보거나 발의 시점 기준이 아니라 현재 시점에서 가진 모든 데이터를 이용해 통과율을 정확도를 높여볼 수 있을 것이다. 또한, 의안 원문에서 의미 정보를 살려 시맨틱(semantic)한 구조를 추출하면 의안 예측의 성능이 더욱 높아질 것을 기대해볼 수 있다.

참고문헌

[1] Poodl: Popong open data library. `http://data.popong.com`, 2014. [Accessed 2014-10-31].

[2] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

[3] Keith T Poole and Howard Rosenthal. *Congress: A political-economic history of roll call voting*. Oxford University Press, 1997.

[4] Tae Yano, Noah A Smith, and John D Wilkerson. Textual predictors of bill survival in congressional committees. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 793–802. Association for Computational Linguistics, 2012.