Sentiment analysis: Bayesian Ensemble Learning

Decision Support Systems 68 (2014) 26–38

김준한, 황근성

서울과학기술대학교 일반대학원 데이터사이언스학과

목 차

1. 서론

- **연구 목적**
  - 언어의 모호성과 관련된 noise sensitivity를 감소 시키기 위해 새로운 앙상블 기법 도입.
    - ✓ 극성 분류 작업의 성능 향상으로 이어짐.
  - 효율적인 후보 앙상블 모델들의 선택 전략 제시 .
  - 리뷰 데이터나 SNS 데이터와 같은 데이터에서 극성 분류에 효과적, 효율적 패러다임 도출.

1. 서론

- **기존에 사용되는 Ensemble의 한계 및 극복**
  - 어떤 모델이 최적의 모델을 나타내는지에 대한 불확실성.
    - ✓ 본 연구의 새로운 앙상블 학습 기법으로 해결.
  - 앙상블 기법은 독립적이고 모두 신뢰성 높은 모델이라는 가정(G.Wang et al,2014)
    - ✓ 실상에선 앙상블 기법들이 독립적이지 않음.
    - ✓ 본 연구에선 각 모델의 기여도를 평가하고, 극성 예측 시 poor classifier를 smooth시켜 개별 classifier들의 의존성과 정확성 설명.
  - 데이터 처리에 드는 시간 및 계산의 복잡성을 무시.
    - ✓ 기존 접근 방식은 계산의 복잡성 문제를 무시하고 가장 높은 성능에 중점
    - ✓ 본 연구에선 계산의 복잡성 문제에 대한 효율적인 방법론을 도출
  - 여러 도메인에 대한 시도 부족.
    - ✓ 전통적인 앙상블 접근법들이 잘 정리된 텍스트(R.Xia et al, 2011)나 noise가 있는 컨텐츠(A.Hassan et al, 2013)에 대한 감성 분석을 시도.
    - ✓ 본 연구에선 짧고 비공식적인 텍스트 (sns 데이터)에서도 제안한 접근법을 시도.
  - 기존 연구들의 한계점을 본 연구에서는 새로운 베이지안 앙상블 기법의 개발로 극복.
    - ✓ 모델의 marginal predictive capability 고려.
    - ✓ backward elimination로 최적의 분류 앙상블 모델 구축.

2. 본론

- **Bayesian Ensemble Learning**
  - 기존의 앙상블 기법들은 개별 모델의 신뢰성을 고려하지 않고 균일한 가중치를 부여.
    - ✓ 베이지안 패러다임을 고려하여 데이터와 모델이 남긴 불확실성 해결.
    - ✓ 가설공간 내 모든 모델들의 marginal prediction capabilities, reliabilities 고려하여 활용가능.

  - Bayesian Model Averaging
    - ✓ 문장 s, classifier들의 집합 C를 고려할 때, 문장 label이 될 확률의 합 (BMA)

$$P(l(s)|C,\mathcal{D}) = \sum_{i \in C} P(l(s)|i,\mathcal{D})P(i|\mathcal{D})$$

    - ✓ Optimal label

$$l^*(s) = \arg\max_{l(s)} P(l(s)|C,\mathcal{D}) = \sum_{i \in C} P(l(s)|i,\mathcal{D})P(i|\mathcal{D})$$
$$= \sum_{i \in C} P(l(s)|i,\mathcal{D})P(\mathcal{D}|i)P(i)$$
$$= \sum_{i \in C} P(l(s)|i,\mathcal{D})P(\mathcal{D}|i).$$

$$P(\mathcal{D}|i) \approx \frac{1}{\iota}\sum_{\iota=1}^{\phi} \frac{2 \times P_{i\iota}(\mathcal{D}) \times R_{i\iota}(\mathcal{D})}{P_{i\iota}(\mathcal{D}) + R_{i\iota}(\mathcal{D})}$$

    - ✓ 앙상블을 구성한 classifier i마다 Φ-fold cross validation을 통해 구한 평균 F1 measure

2. 본론

- **Bayesian Ensemble Learning**
  - Model selection strategy
  - 앙상블의 중요한 문제는 Ensemble에 포함될 최적의 Model set을 선택하는 것.
    - ✓ classifier들의 공헌도를 평가 해야함.
    - ✓ 본 연구에서는 주어진 모델 앙상블의 각 classifier가 제공하는 discriminative marginal contribution를 계산할 수 있는 heuristic 접근법을 제안.
    - ✓ classifier i의 공헌도

$$r_i^S = \frac{\sum_{j\in\{S\setminus i\}}\sum_{q\in\{0,1\}} P(i=1|j=q)P(j=q)}{\sum_{j\in\{S\setminus i\}}\sum_{q\in\{0,1\}} P(i=0|j=q)P(j=q)}$$

  - backward elimination
    - ✓ 전체 집합에서 관련 모델을 제거하면 평가 값이 감소할 수 있으며, 불완전한 집합에 관련 모델을 추가하는 것은 즉각적인 영향을 미침.
    - ✓ Full-set 에서 가장 낮은 공헌도 r_i^S 가 나온 classifier를 제거하고 반복적으로 r_i^S 계산 → 최적의 r_i^S 가 나오는 조합 선택.

step	SGD	Logistic	MNB	RF	Ridge	ACC
1	1.7655	1.9730	2.1915	**1.0900**	1.6557	**1.7351**
2	1.5078	1.7020	1.8984	-	**1.4242**	1.6332
3	**1.4949**	1.5695	1.7787	-	-	1.6144
4	-	**1.4240**	1.6627	-	-	1.5433

* Dataset : MovieData

2. 본론

- **Experimental investigation**
  - Weighting schema
    - ✓ 전반적으로 text 문장에 같은 용어가 다중으로 존재하면 이것을 표현하기 힘듦
    - ✓ 본 연구에서 각 문장은 가중치 w를 계산할 수 용어로 구성된 벡터로 표현
    - ✓ 가중치 w를 계산하기 위해 서로 다른 weight scheme를 사용
  - Weighting schema 종류
    - ✓ Boolean: f(t,d) = t (t가 d에 한 번이라도 나타나면 1, 아니면 0)
    - ✓ Term Frequency(TF):

$$\mathrm{tf}(t,d) = 0.5 + \frac{0.5 \times \mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$

(문서-d, 단어-t, 문서 d내에 단어 t의 총 빈도-tf(t,d))

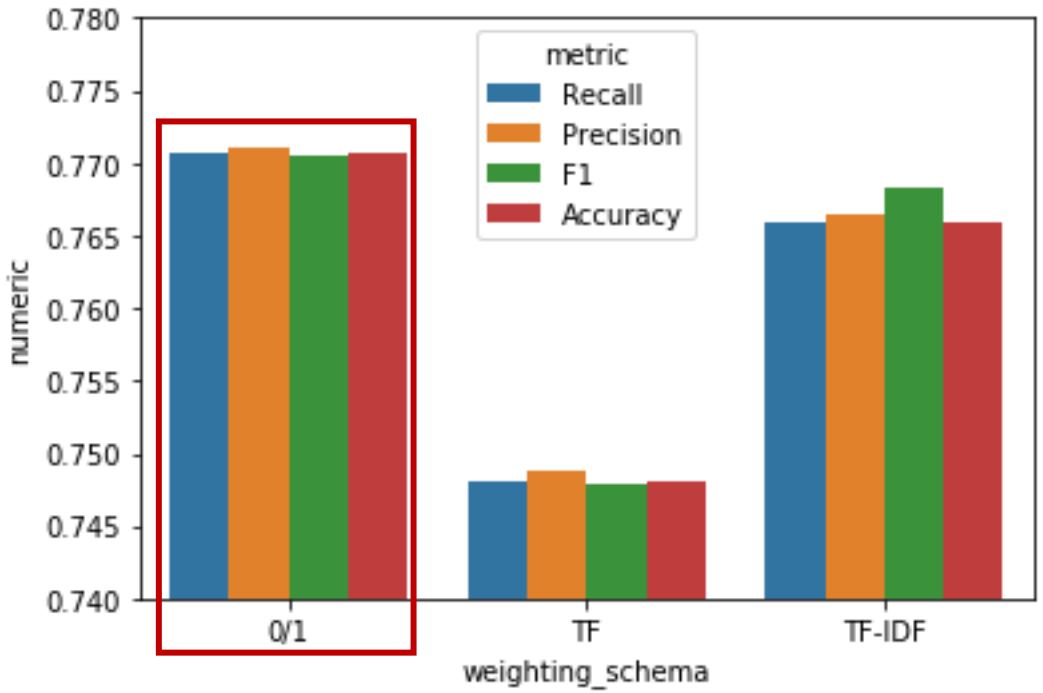
    - ✓ Term Frequency Inverse Document Frequency(TF-IDF):

$$\mathrm{tf}(t,d) \times \mathrm{idf}(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

(idf(t,D)-한 단어가 전체 문서에서 얼마나 공통적으로 나타나는지, |D|-전체 문서의 수, |{d ∈ t : t ∈ d}|-단어 t가 포함된 문서의 수)

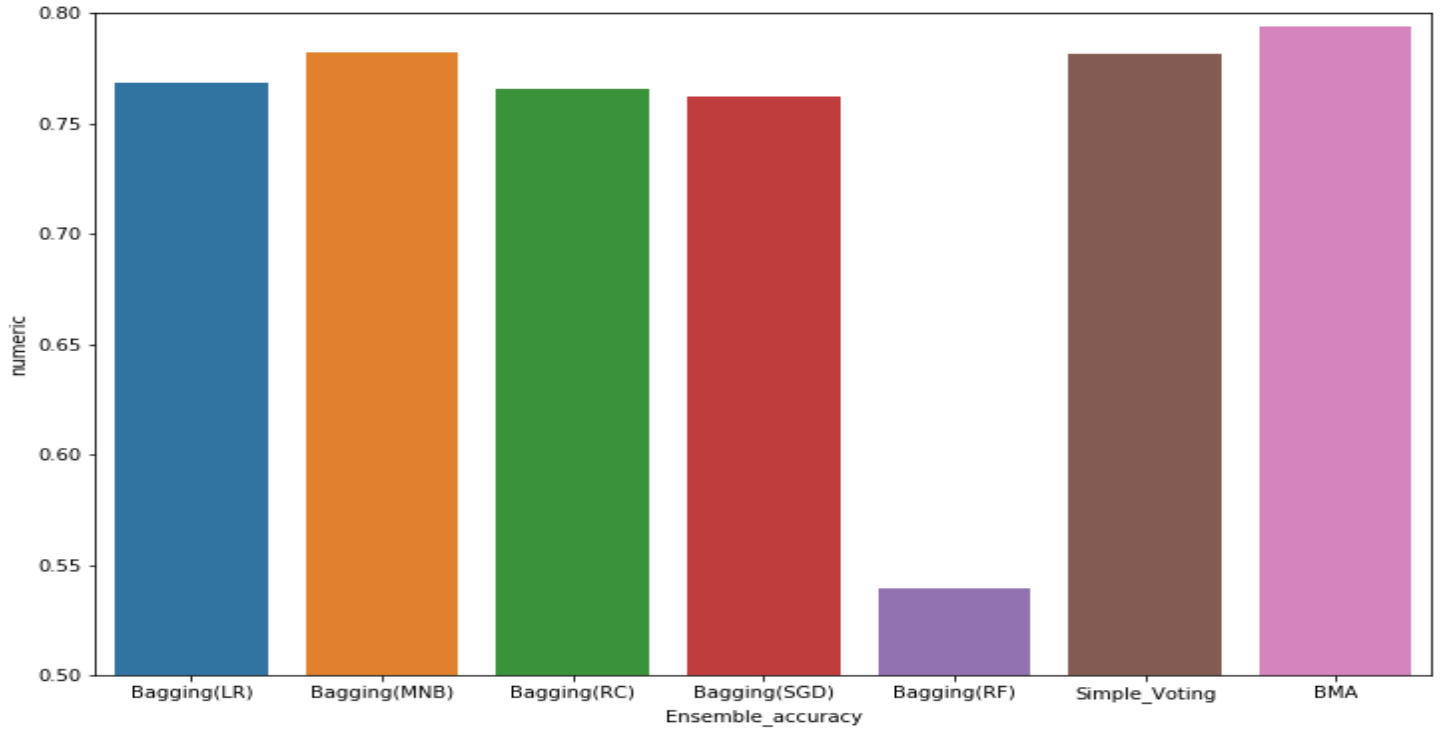
2. 본론

- **Computational results**
  - weighting schema(dataset-MovieData)
    - ✓ 최적의 weighting schema -> 제일 좋은 성능인 Boolean(0/1) 선택



2. 본론

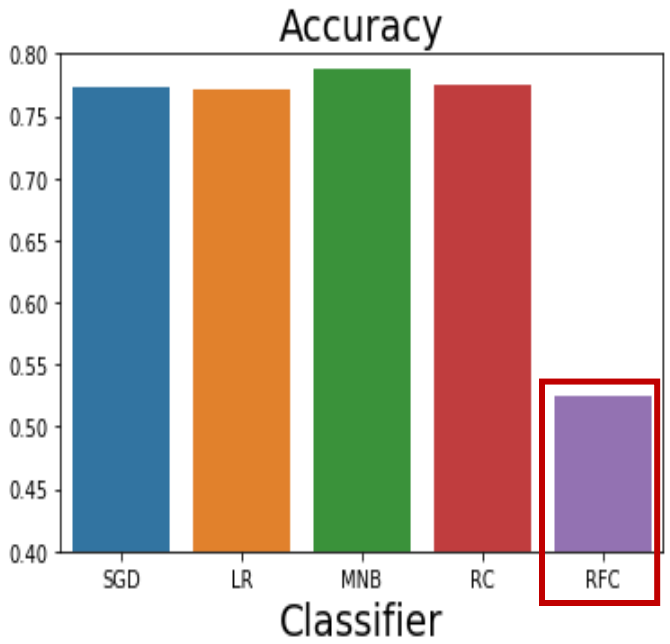
- **Computational results**
  - Baseline and ensemble classifiers
    - ✓ baseline classifier: Linear Regression(LR), Multinomial Naive Bayes(MNB), SGD classifier(SVM), Ridge Classifier(RC), RandomForestClassifier(RF)
    - ✓ Ensemble: Simple Voting, Bagging



    - ✓ Ensemble마다 비교했을 때, bagging과 Simple Voting의 성능이 일부를 제외하고 비슷하게 나옴
    - ✓ BMA는 모든 Ensemble과 비교하여 성능이 더 좋게 나옴.

2. 본론

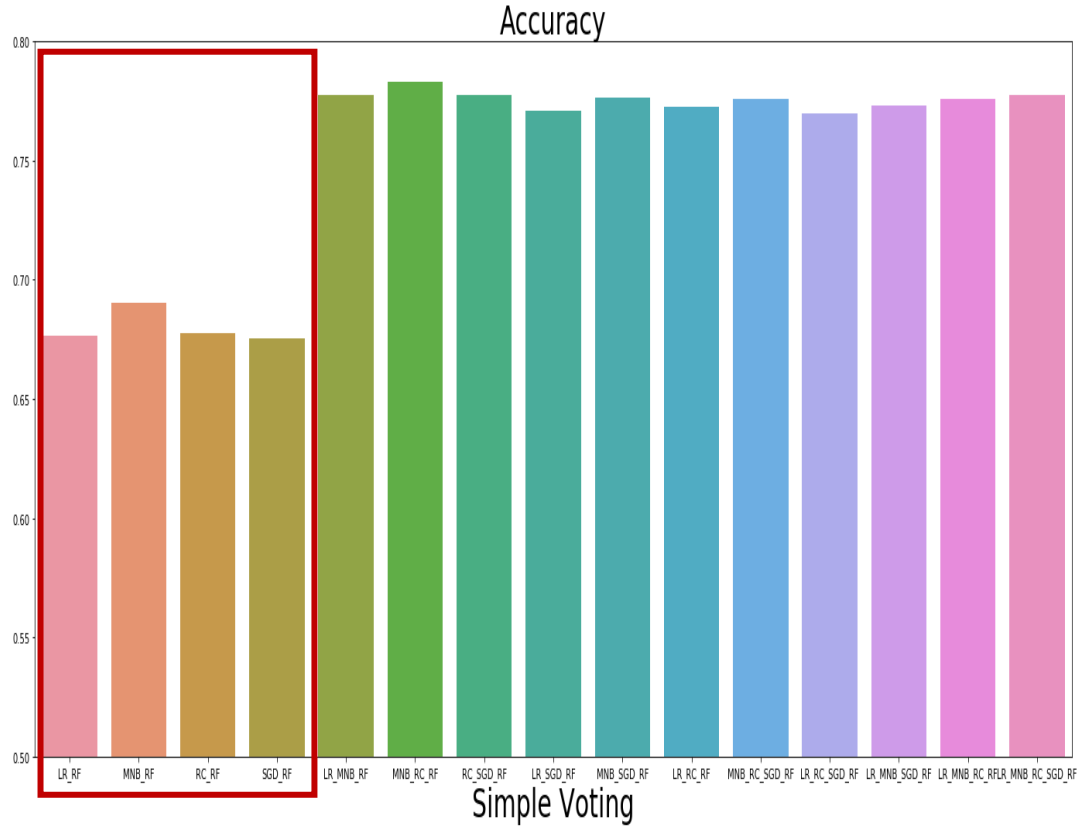
- **Computational results**
  - baseline classifier와 SV, BMA Accuracy 비교

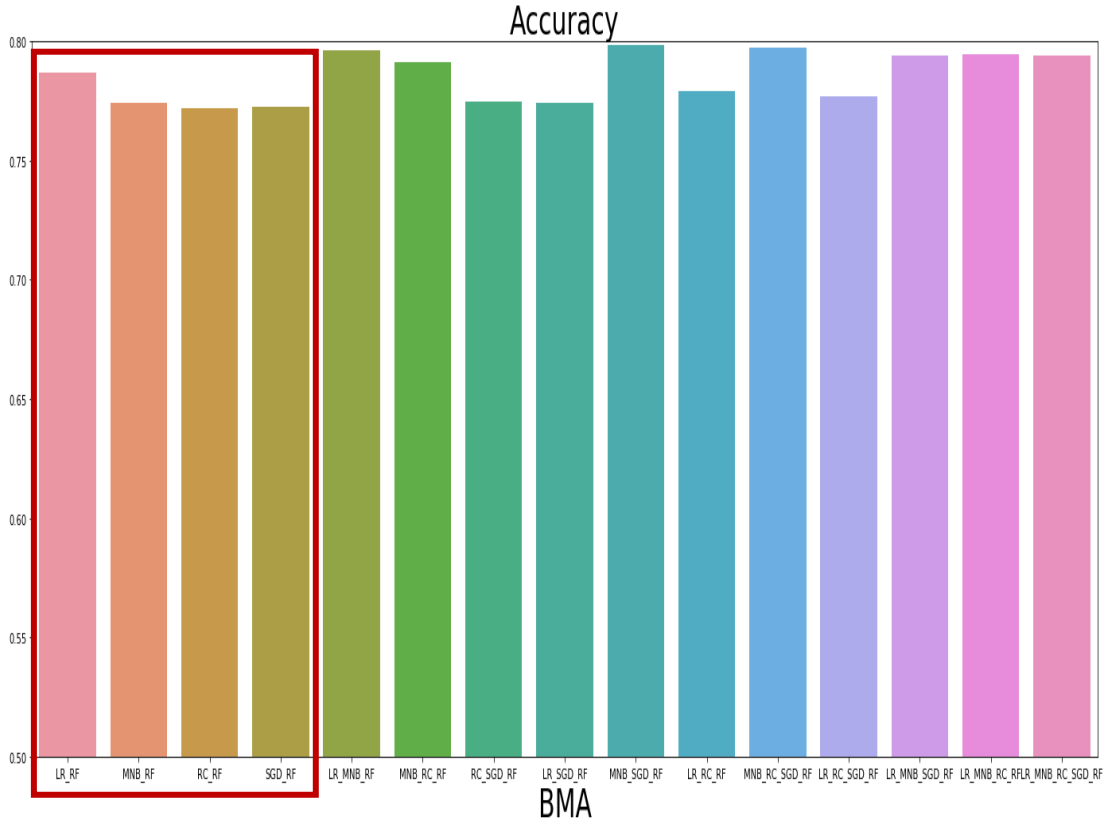


✓ 개별 classifier의 accuracy만 놓고 볼 때, Random Forest의 accuracy가 가장 낮게 나옴.

2. 본론

- **Computational results**
  - baseline classifier와 SV, BMA Accuracy 비교





✓ 성능이 낮게 나온 Random Forest를 기준으로 하여 Ensemble set을 구성함.

✓ 대부분의 BMA가 SV의 accuracy보다 높게 측정됨.

3. 결론

- **Conclusion**
  - 구현의 한계점
    - ✓ 본 연구에 쓰이던 Dataset이 없는 경우가 많아 Computational complexity analysis 구현에 제약을 받음.
    - ✓ Ensemble combination rules인 MV, MEAN, MAX, PRODUCT를 사용하지 못하여 RI를 통한 Bagging 비교를 수행하지 못함.
  - 구현의 의의
    - ✓ BMA를 통해 서로 다른 classifier의 전략적인 조합을 정의하는 능력이 다른 앙상블 기법에 비해 효율적임을 보임.
  - 향후 연구방향
    - ✓ Binary Classification 이외에 MultiClass Classification 문제에도 해당 연구의 기법을 적용할 수 있음.