랜덤포레스트를 이용한 국내 학술지 논문의 자동분류에 관한 연구

An Analytical Study on Automatic Classification of Domestic Journal articles Using Random Forest

김판준 (Pan Jun Kim)*

초 록

대표적인 앙상블 기법으로서 랜덤포레스트(RF)를 문헌정보학 분야의 학술지 논문에 대한 자동분류에 적용하였다. 특히, 국내 학술지 논문에 주제 범주를 자동 할당하는 분류 성능 측면에서 트리 수, 자질선정, 학습집합 크기 등 주요 요소들에 대한 다각적인 실험을 수행하였다. 이를 통해, 실제 환경의 불균형 데이터세트(imbalanced dataset)에 대하여 랜덤포레스트(RF)의 성능을 최적화할 수 있는 방안을 모색하였다. 결과적으로 국내 학술지 논문의 자동분류에서 랜덤포레스트(RF)는 트리 수 구간 100~1000(C)과 카이제곱통계량(CHI)으로 선정한 소규모의 자질집합(10%), 대부분의 학습집합(9~10년)을 사용하는 경우에 가장 좋은 분류 성능을 기대할 수 있는 것으로 나타났다.

ABSTRACT

Random Forest (RF), a representative ensemble technique, was applied to automatic classification of journal articles in the field of library and information science. Especially, I performed various experiments on the main factors such as tree number, feature selection, and learning set size in terms of classification performance that automatically assigns class labels to domestic journals. Through this, I explored ways to optimize the performance of random forests (RF) for imbalanced datasets in real environments. Consequently, for the automatic classification of domestic journal articles, Random Forest (RF) can be expected to have the best classification performance when using tree number interval 100~1000(C), small feature set (10%) based on chi-square statistic (CHI), and most learning sets (9-10 years).

키워드: 자동분류, 자동주석, 디지털 큐레이션, 학술지 논문, 랜덤포레스트(RF), 복수-범주 분류, 불균형 데이터, 자질선정

automatic classification, automatic annotation, digital curation, journal articles, random forest (RF), multi-label classification, imbalanced data, feature selection

* 신라대학교 문헌정보학과 부교수(pjkim@silla.ac.kr)

■ 논문접수일자 : 2019년 5월 15일 ■ 최초심사일자 : 2019년 6월 21일 ■ 게재확정일자 : 2019년 6월 21일

■ 정보관리학회지, 36(2), 57-77, 2019. 〔http://dx.doi.org/10.3743/KOSIM.2019.36.2.057〕

1. 서 론

전 세계적으로 학술정보의 생산 및 유통이 폭발적으로 증가하는 가운데, 기계학습 알고리즘에 기초한 디지털 큐레이션의 필요성이 나날이 증가하고 있다. 디지털 큐레이션은 디지털 정보의 수집(acquisition), 선정(selection), 주석(annotation), 보존(maintenance)을 포괄하는 것으로, 이 중에서 특히 디지털 정보에 대한 주석은 전통적인 사서의 핵심 업무인 분류(또는 색인)에 해당한다(Lok, 2010; Ma, 2017; Ma, Zhang, Sunderraman, Fox, Laird, Turner, & Turner, 2015; Turner, Chakrabarti, Jones, Xu, Fox, Luger, Laird, & Turner, 2013; Trieschnigg, Pezik, Lee, Jong, Kraaij, & Rebholz-Schuhmann, 2009). 그러나 학술지 논문에 대한 수작업 분류 또는 주석(manual classification or manual annotation)은 막대한 시간과 노력을 필요로 하는 까닭에 전문가(curators)에 의한 고품질의 분류 데이터가 턱없이 부족한 상황이며, 연구자는 자신의 논문에 대한 분류를 수행하기 위한 전문지식과 기술이 부족함은 물론 관심 자체가 거의 없는 것이 현실이다(Brandenburg, 2017; Ma, 2017). 특히 학술지 논문에 대한 분류 정보가 거의 제공되지 않고 있는 국내 데이터베이스 환경에서, 폭발적으로 증가하고 있는 학술지 논문에 대한 수작업 분류는 현실적으로 거의 불가능하다(김판준, 2018; 김판준, 2016). 따라서 전문가에 의한 수작업 분류를 지원할 수 있는 효과적인 방법으로 기계학습 알고리즘에 기초한 자동분류를 적극적으로 모색할 필요가 있다.

기존의 기계학습 기반 자동분류는 대부분 범주의 분포가 균등한 데이터(balanced data)를 전제로 하고 있다(Wu, Ye, Zhang, Ng, & Ho, 2014). 그러나 실제 환경에서 학술지 논문의 분류는 하나의 논문에 다수의 범주가 할당되는 복수-범주 분류에 해당하며, 그 결과로 할당된 범주 정보는 범주별 문헌 분포가 균등하지 않은 불균형 데이터(imbalanced data)인 경우가 대부분이다(김판준, 2018; Madjarov, Kocev, Gjorgjevikj, & Džeroski, 2012). 이러한 데이터 불균형 문제는 대부분의 학습문헌이 포함되는 대범주 또는 학습문헌이 거의 없는 소범주들이 존재함은 물론 하나의 학습문헌이 복수의 범주에 동시에 속할 수 있기 때문에 발생하며, 분류 알고리즘의 성능을 저하시키는 주요 원인 중 하나가 된다(Kim, Kang, & Kim, 2015; Ma et al., 2015; Nayak, Ramesh, & Shah, 2013; Turner et al., 2013; Wu et al., 2014).

대표적인 앙상블 분류 알고리즘으로서 최근 다양한 분야에 활발하게 적용되고 있는 랜덤포레스트(Random Forest: RF)[1]는 텍스트 처리에서 여러 장점을 가지고 있으며, 불균형한 이진 분류에서 우수한 성능을 보인 것으로 보고되었다(Brown & Mues, 2012; Xu, Guo, Ye, & Cheng, 2012; Yao, Yang, & Zhan, 2013). 또한 설명변수가 다수일 때 예측력이 매우 높고 매우 안정적인 모형을 제공하면서(Siroky, 2009), 과적합(overfitting)에 강건하고 잡음(noise)이나 이상치(outlier)의 영향이 적은 장점을 갖고 있다. 따라서 랜덤포레스트(RF)는

1) 이후 랜덤포레스트(RF)로 표기함.

텍스트 데이터이면서 고차원 자질집합과 불균형한 범주집합으로 구성된 국내 학술지 논문의 자동분류에 적합한 기법이라 할 수 있다(김성진, 안현철, 2016; Dogan & Uysal, 2018; Yao, Yang, & Zhan, 2013).

본 연구는 국내 학술지 논문에 적절한 주제범주를 자동 할당하는 목적으로 랜덤포레스트(RF)를 적용하였다. 특히, 분류 성능 측면에서 트리 수, 자질선정, 학습집합 등 주요 요소들에 대한 다각적인 실험을 수행하고, 그 결과를 분석하여 실제 환경의 불균형 데이터세트인 국내 학술지 논문의 자동분류에 가장 적절한 랜덤포레스트(RF)의 적용 방안을 제시하였다.

2. 이론적 배경

2.1 랜덤포레스트(RF)

랜덤포레스트(RF)는 무작위로 선택된 데이터 하부집합(subsets)과 자질집합(feature sets)으로 학습시킨 의사결정 트리에 기반한 대표적인 앙상블 분류 알고리즘으로 Breiman(2001)이 개발하였다. 각 의사결정 트리는 데이터의 부트스트랩 샘플로 구성되며, 노드 분할은 무작위로 선택된 자질집합 중에서 가장 좋은 자질에 기초한다. 개별 결정트리는 입력 벡터의 범주를 독립적으로 할당하고 최종적인 분류 범주는 각 분류기의 결과를 조합(voting or averaging)하여 결정된다(Dogan & Uysal, 2018; Fawagreh, Gaber, & Elyan, 2014). 기존에 연구되어온 의사결정나무에서는 모든 변수를 사용하여 가장 최적의 결과를 내는 분할로 각각의 노드(node)를 나타낸 것과 달리, 랜덤 포레스트에서는 각각의 노드를 나타낼 때 설명변수를 무작위로 선택하고 선택된 설명변수의 집합 중에서 가장 최적의 결과를 내는 방법을 이용한다(권안나, 2013). 즉, 학습집합에서 복원 추출에 의해 부스트랩 데이터를 생성을 N번 반복하여 N개의 부스트랩 데이터를 생성하고, 의사결정나무 알고리즘을 적용할 때 각각의 노드에서 랜덤하게 m개의 설명변수를 선택하는 것이다.

이러한 랜덤포레스트(RF)는 학습과정에서 배깅(bagging)과 임의노드 최적화(randomized node optimization)를 통해 노이즈와 과적합에 강건하며, 복수의 분류기로 구성되기 때문에 개별 분류기보다 정확한 분류가 가능하다(Boinee, Angelis, & Foresti, 2005; Brandenburg, 2017). 또한, 고차원 벡터로 구성되는 텍스트의 분류에 좋은 성능을 보이며 매우 안정적인 모형을 제공한다(Siroky, 2009).

이에 따라 랜덤포레스트(RF)는 기계학습 분야에서 많은 관심과 연구가 집중되고 있으며 분류, 예측, 자질선정, 불균형 데이터 등 다양한 목적으로 응용되고 있다. 해외에서는 2000년대 초반부터 생물정보학과 의학, 생태학 분야에서 랜덤포레스트(RF)를 적용한 많은 연구가 이루어졌다(Austin, Tu, Ho, Levy, & Lee, 2013; Cutler, Edwards, Beard, Cutler, Hess, Gibson, & Lawler, 2007; Ward, Pajevic, Dreyfuss, & Malley, 2006). 이외에도 천문학(Gao, Zhang, & Zhao, 2009), 농학(Löw, Schorcht, Michel, Dech, & Conrad, 2012), 범죄학(Berk, Li, & Hickman, 2005) 등 다양한 분야에서 랜덤포레스트(RF)를 적용한 연구가 수행되었다. 특히, 뉴스기사, 웹문서, 사건 보고서, 학술지 논문, 종

교 경전, 이메일 등 다양한 텍스트 데이터를 대상으로 랜덤포레스트(RF)를 적용한 연구도 활발히 진행되고 있다(Afianto & Adiwijaya, 2017; Aung, Myanmar, & Hla, 2009; Brandenburg, 2017; Klassen & Paturi, 2010; Liparas, HaCohen-Kerner, Moumtzidou, Vrochidis, & Kompatsiaris, 2014; Ma, 2017; Wu et al., 2014).

국내에서도 2000년대에 들어서서 의학 분야를 중심으로 랜덤포레스트(RF) 관련 연구가 활발하게 진행되었다(윤태관, 이관수, 2008; 이현주, 신동규, 박희원, 김수한, 신동일, 2011). 또한 공학 분야(정준호, 장경현, 김재협, 2016; 최혁진, 최성욱, 한경숙, 2012; 홍준혁, 고병철, 남재열, 2013)는 물론 통계학(권안나, 2013), 경제학(김성진, 안현철, 2016; 서종덕, 2016)에서도 다양한 목적으로 랜덤포레스트(RF)를 적용하였다. 특히, 국내에서 텍스트 데이터를 대상으로 랜덤포레스트(RF)를 적용한 것으로는 한국어의 상호참조 문제(정석원, 최맹식, 김학수, 2016)와 국내 드라마의 등급 분류(강수연, 전희정, 김지혜, 송정우, 2015), 청소년의 진로 선택 여부(유진은, 2015), 신문사 보도 특성 분석(조현채, 박철용, 2018)에 관한 연구들이 있다.

2.2 랜덤포레스트(RF)의 성능 요소

랜덤포레스트(RF)의 분류 성능에 영향을 미치는 요소는 다양하지만, 지금까지 선행연구에서 주로 다루어진 대표적인 성능 요소는 트리 수, 자질 가중치, 자질선정, 학습집합 크기 등이다(Ma & Fan, 2017). 특히, 랜덤포레스트(RF)의 성능 향상을 위한 요소로는 트리 수(Choi & Kim, 2016; Latinne, Debeir, & Decaestecker, 2001), 자질 가중치(홍준혁, 고병철, 남재열, 2013; Xu et al., 2012; Xu, Huang, Williams, & Ye, 2012), 결과 통합 방법(Fawagreh, Gaber, & Elyan, 2014; Robnik-Šikonja, 2004; Tsymbal, Pechenizkiy, & Cunningham, 2006)에 관한 연구가 활발하였다.

모든 단어를 분류 자질로 이용하는 것보다 문헌의 내용을 대표할 수 있는 단어를 사용할 때 분류 성능이 향상된다는 것은 선행연구를 통하여 입증된 사실이다(김판준, 2016; Yang & Pedersen, 1997). 따라서 랜덤포레스트(RF)의 성능 향상을 위한 방법으로 자질선정을 적용한 연구가 다양한 측면에서 이루어졌다(Amaratunga, Cabrera, & Lee, 2008; Dogan, & Uysal, 2018; Kong, Gong, Ding, & Hou, 2017; Ma, 2017; Zhou, Zhou, & Li, 2016). 또한, 다른 분류 알고리즘과 랜덤포레스트(RF)의 성능을 비교한 것으로는 도시 토지 표지 분류에 자질선정을 적용하여 3개 분류기(나이브베이즈(NB), 지지벡터기계(SVM), 랜덤포레스트(RF))의 성능을 비교한 연구(이진욱, 유국현, 문병민, 배석주, 2017; Dogan & Uysal, 2018)와 최근 주목을 받고 있는 딥러닝(Deep Learning)과 랜덤포레스트(RF)의 성능을 비교한 연구가 수행되었다(남승현, 오명섭, 김성관, 강창완, 김규곤, 최승배, 2017).

그러나 지금까지 국내에서 학술지 논문의 분류에 랜덤포레스트(RF)를 적용한 연구는 찾아볼 수 없으며, 해외에서도 학술지 논문의 분류 목적으로 랜덤포레스트(RF)의 주요 성능 요소를 다각적으로 살펴본 연구는 찾아보기 어렵다(Ma & Fan, 2017). 따라서 본 연구에서는 국내 학술지 논문의 자동분류에 랜덤포레스트(RF)를

적용하는 목적으로 주요 성능 요소들에 대한 다각적인 실험을 수행한 결과로서, 실제 환경의 텍스트 데이터인 국내 학술지 논문의 분류에 가장 적합한 성능 요소의 적용 방안을 모색하였다.

3. 연구 방법

3.1 연구 문제

실제 환경의 불균형 데이터인 국내 학술지 논문집합에 두 가지 범주 할당 방법(단일-범주, 복수-범주)으로 범주를 자동 할당하는 환경에서, 랜덤포레스트(RF)의 성능에 영향을 미치는 주요 요소들에 대하여 연구문제를 설정하였다. 지금까지 의사결정 트리에 기초한 랜덤포레스트(RF) 관련 연구에서 주로 다루어진 성능 요소는 트리 수, 자질선정, 학습집합 크기 등이다. 이에 따라 본 연구는 랜덤포레스트(RF)의 성능에 영향을 미치는 주요 요소들에 대한 연구문제를 다음과 같이 크게 세 가지로 설정하였다.

- 연구문제 1. 국내 학술지 논문의 자동분류에서 랜덤포레스트(RF)는 트리 수의 변화에 따른 분류 성능에 차이가 있는가?
- 연구문제 2. 국내 학술지 논문의 자동분류에서 랜덤포레스트(RF)는 자질선정 방법과 자질집합에 따른 분류 성능에 차이가 있는가?
- 연구문제 3. 국내 학술지 논문의 자동분류에서 랜덤포레스트(RF)는 학습집합의 크기에 따른 분류 성능에 차이가 있는가?

3.2 실험

3.2.1 실험 환경

본 연구의 실험 문헌집단은 문헌정보학 분야의 『정보관리학회지』에 수록된 최근 14년(2002년~2015년)의 논문 중에서, 한글로 작성되고 저자 키워드와 초록이 있는 문헌정보학 분야의 논문 651편이다. 이중에서 이전 10년(2002년~2011년)의 453편(70%)을 학습집합, 이후 4년(2012년~2015년)의 198편(30%)은 검증집합으로 구성하였다. 또한, 랜덤포레스트(RF)를 적용한 자동분류 실험에 사용된 문헌집합, 자질집합, 자질 가중치, 범주집합은 다른 분류 알고리즘을 적용한 저자의 선행연구의 환경과 동일하게 구성하였다. 따라서 본 연구의 문헌집합은 기존의 실험 문헌집합(Reuters-21578, 20-newsgroups, OHSUMED)에 비해 대부분 학습문헌의 수가 상대적으로 적은 중저빈도 범주로 구성되어 있으며, 각 범주 당 학습문헌 수의 편차가 큰 불균형 데이터(imbalanced data)에 해당한다(김판준, 2018).

실험에 사용된 랜덤포레스트(RF) 알고리즘은 파이썬에서 제공하는 scikit-learn 라이브러리의 RandomForestClassifier 모듈을 사용하여 구현하였다. 기본적으로 배깅(bagging; bootstrap aggregating)을 통해 구성한 학습문헌 집합으로 생성한 트리를 학습하였고, 불순도(impurity) 측정에는 지니계수(Gini index)를 사용하였다. 이외에 연구 문제로 설정한 세 가지 성능 요소를 제외한 랜덤포레스트(RF) 관련 파라미터는 모듈에서 제공하는 기본 설정 값(default value)을 그대로 사용하였다.

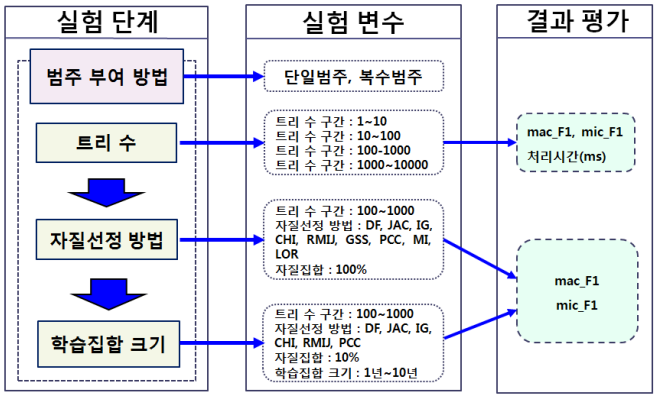
지금까지 자동분류의 성능 척도로 많이 사용

되어 온 산술평균 정확도(accuracy)는 대범주의 성능에 크게 영향을 받는 문제가 있으므로(김판준, 2018; Kim, Kang, & Kim, 2015), 범주별 편차가 큰 불균형 데이터인 본 연구의 실험 문헌집합에 대한 평가 척도로 부적합한 측면이 있다. 이에 따라 본 연구에서는 성능 평가 척도로 매크로 평균 F1(mac_F1)과 마이크로 평균 F1(mic_F1)[2]을 함께 사용하였다. 자동분류의 성능 평가에서 실제 환경의 불균형 데이터를 대상으로 자동분류의 성능에 영향을 주는 세 가지 주요 성능 요소를 다각적으로 검토하기 위하여, 서로 다른 특성을 가진 mac_F1과 mic_F1을 함께 산출하였다. 특히, 복수-범주의 성능 평가는 기존의 척도에 완화된 기준을 적용하여 조정한 복수-범주 mac_F1과 복수-범주 mic_F1을 사용하였다(김판준, 2018).

랜덤포레스트(RF)는 각각의 랜덤포레스트 과정을 N번 반복하여 분류를 수행한 결과의 평균으로 분류 성능을 산출하는데, 선행연구에서는 반복 횟수(N=10, 30, 50, 100)를 임의로 정하여 적용하고 있다(남승현 외, 2017; 정준호, 장경현, 김재협, 2016; Choi & Kim, 2016; Ma & Fan, 2017; Yao, Yang, & Zhan, 2013). 본 연구에서는 사전실험에서 트리 수를 1개~10000개까지 증가시키면서 반복 횟수(N)를 10회와 50회로 적용한 성능이 동일하거나 거의 차이가 없었기 때문에, 랜덤포레스트(RF)의 반복 횟수(N)를 10회로 하여 분류 성능을 산출하였다.

3.2.2 실험 단계

랜덤포레스트(RF)의 성능에 영향을 미치는 세 가지 주요 요소별로 연구 문제를 설정하고, 국내 학술지 논문에 대한 자동분류의 성능을 비교하는 실험을 수행하였다. 실험 문헌집단을 학습집합(10년, 70%)과 검증집합(4년, 30%)으로 구분한 다음, 학습문헌에 부여된 범주를 학습하여 이후의 입력문헌으로서 검증문헌에 자동 할당한 결과를 수작업 분류 결과와 비교하여 성능을 평가하였다. 랜덤포레스트(RF)를 적용한 자동분류 실험 단계별 변수와 결과의 평가 방법은 〈그림 1〉과 같다.



〈그림 1〉 실험 단계별 변수와 평가 방법

2) 이후 mac_F1, mic_F1으로 표기함.

4. 실험 결과 및 분석

4.1 트리 수

랜덤포레스트(RF)에서 일반적으로 트리 수의 증가는 개별 분류기의 다양성을 보장하여 성능을 향상시키는 것으로 알려져 있다. 그러나 트리 수가 증가할수록 시간 비용이 증가하고 해석 가능한 결과가 줄어들 수 있는 반면, 트리 수가 감소하면 분류 오류가 증가하고 성능이 하락하는 문제가 있다(Ma & Fan, 2017). 따라서 분류 성능과 처리시간 측면에서 최적의 트리 수를 찾기 위한 실험을 수행하였다. 먼저, 랜덤포레스트(RF) 트리 수를 4개 구간(1~10, 10~100, 100~1000, 1000~10000)으로 구분하고, 두 가지 범주 할당 방법(단일-범주, 복수-범주)에 따라 분류한 결과를 mac_F1, mic_F1으로 산출한 결과는 〈표 1〉과 같다.

〈표 1〉에서 트리 수가 늘어날수록 범주 할당 방법에 상관없이 구간별 평균 성능이 상승하는 것을 알 수 있다. 그러나 트리 수 구간이 1~10(A)에서 10~100(B), 그리고 10~100(B)에서 100~1000(C)으로 증가한 경우에 성능이 크게 향상되는 것에 비하여, 100~1000(C)에서 1000~10000(D)으로 증가한 경우의 성능 차이는 크지 않았다.

〈그림 2〉는 랜덤포레스트(RF) 분류의 처리시간[3] 측면에서 트리 수 구간별 성능을 살펴본 결과이다. 각 트리 수 구간별로 평균학습시간과 평균검증시간을 측정한 결과에 따르면, 1~10(A)에서 10~100(B), 그리고 10~100(B)에서 100~1000(C)으로 증가하는 경우에는 범주 할당 방법에 상관없이 처리시간에 크게 차이가 없었다. 그러나 트리 수 구간을 1000~10000(D)으로 늘린 경우에는 평균학습시간이 급격히 증가하였으며, 특히 복수-범주 평균학습시간은 100~1000(C)보다 무려 10배 이상의 시간이 소요되었다. 이에 따라 분류 성능(mac_F1, mic_F1)과 처리시간(ms) 측면에서 가장 적절한 트리 수가 100~1000(C) 구간인 것으로 판단하였고, 이후의 실험에서는 트리 수를 100~1000(C) 구간으로 설정하여 진행하였다.

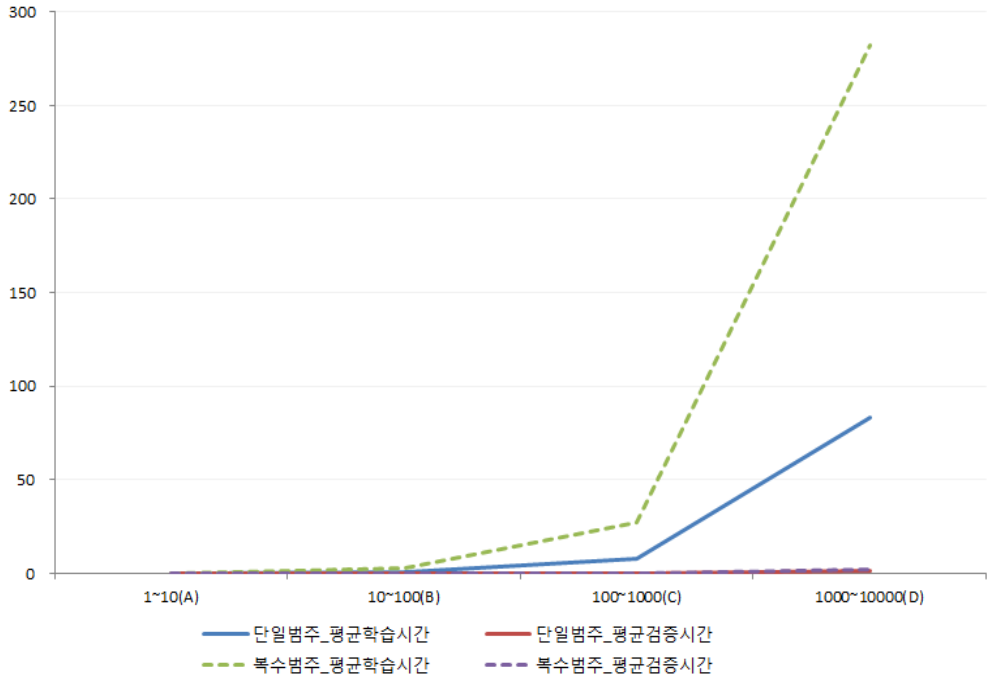
4.2 자질선정

랜덤포레스트(RF)에 선행연구에서 많이 사용된 자질선정 방법을 적용한 분류 성능을 살펴보았다. 특히, 지금까지 자동분류 연구에서

〈표 1〉 랜덤포레스트(RF)의 트리 수 구간별 성능: mac_F1, mic_F1

구분	트리 수 구간별 평균 성능				성능 차이 (D-C)
	1~10(A)	10~100(B)	100~1000(C)	1000~10000(D)	
단일-범주 mac_F1	0.3249	0.4558	**0.4923**	**0.5008**	**0.0085**
단일-범주 mic_F1	0.4347	0.5736	**0.6102**	**0.6169**	**0.0067**
복수-범주 mac_F1	0.4993	0.6674	**0.6775**	**0.6833**	**0.0058**
복수-범주 mic_F1	0.5538	0.7017	**0.7247**	**0.7305**	**0.0058**

3) 처리 시간의 단위는 밀리세컨드(milliseconds, ms)로 1초를 1000으로 나눈 값이다.



〈그림 2〉 랜덤포레스트(RF)의 트리 수 구간별 성능: 처리 시간(단위: ms)

많이 사용된 자질선정 방법 9개(문헌빈도/DF, 자카드 계수/JAC, 정보획득량/IG, 카이제곱통계량/CHI, 상대적 상호정보량J/RMIJ, GSS 계수/GSS, 피어슨 계수/PCC, 상호정보량/MI, 로그승산비/LOR)를 사용하였다(김판준, 2006; 이재윤, 2005; Lee & Kim, 2015; Manning & Raghavan, 2008; Roul & Rai, 2016; Yang & Pedersen, 1997). 〈표 2〉는 이러한 자질선정 방법 9개를 적용하여 자질집합을 증가하는 경우에 단일-범주 할당 방법에 대한 랜덤포레스트(RF)의 분류 성능을 mac_F1으로 산출한 것이다. 여기서 각 성능 수치는 이전 단계 실험에서 정한 트리 수 100~1000(C) 구간과 각 분류 과정을 10회 반복한 평균값으로 산출하였다.

〈표 2〉에서 랜덤포레스트(RF)는 단일-범주를 부여하는 경우에 자질선정 기법 중에서 카이제곱통계량(CHI)이 mac_F1 측면에서 가장 좋은 성능을 보였다. 특히, 카이제곱통계량(CHI)은 가장 작은 자질집합(10%)에서 모든 자질선정 기법 가운데 최고 성능이었으며(0.5189), 비교적 소규모의 자질집합을 사용하여도 지속적으로 좋은 성능을 보였다. 반면, 다른 자질선정 기법들은 최소 40% 이상(PCC, RMIJ, DF)의 자질집합을 사용하거나 90% 이상(MI)의 자질집합을 사용하여야 좋은 성능을 기대할 수 있는 것으로 나타났다. 상대적 상호정보량J(RMIJ), 피어슨 계수(PCC), 상호정보량(MI)은 카이제곱통계량(CHI)보다 더 많은 자질집합을 사용하는 경우에도 상대적으로 낮은 성능을 보였다.

〈표 3〉은 9개 자질선정 기법별로 자질집합을 증가하는 경우에 단일-범주 할당 방법에 대한 랜덤포레스트(RF)의 분류 성능을 mic_F1

〈표 2〉 자질선정을 적용한 랜덤포레스트(RF) 분류 성능: 단일-범주, mac_F1

구분	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
DF	0.4696	0.4928	0.4989	0.5022	0.5056	**0.5094**	0.5051	**0.5050**	0.5010	0.4948
JAC	0.4730	0.4930	0.5009	0.5014	0.5052	0.5027	0.5001	0.5001	0.4977	0.4971
IG	0.4846	0.4927	0.4942	0.4991	0.4978	0.4987	0.4910	0.4961	0.4954	0.4955
CHI	**0.5189**	**0.5048**	**0.5082**	0.5073	0.5044	0.5000	0.4991	0.4982	0.5006	0.4978
RMIJ	0.4288	0.4336	0.5042	0.5054	**0.5067**	0.5084	**0.5072**	0.5031	0.5008	0.4970
GSS	0.3899	0.4282	0.4327	0.4745	0.4639	0.4549	0.4537	0.4386	0.4272	0.4978
PCC	0.4531	0.4219	0.4951	**0.5122**	0.4668	0.4461	0.4227	0.4265	0.4377	0.4965
MI	0.1537	0.1651	0.3003	0.2460	0.4397	0.4919	0.4961	0.4745	**0.5084**	**0.4987**
LOR	0.2640	0.3421	0.4163	0.4285	0.4304	0.4189	0.4174	0.4091	0.4060	0.4981
max	**0.5189**	0.5048	0.5082	**0.5122**	**0.5067**	**0.5094**	**0.5072**	0.5050	**0.5084**	**0.4987**

〈표 3〉 자질선정을 적용한 랜덤포레스트(RF) 분류 성능: 단일-범주, mic_F1

구분	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
DF	0.6135	0.6318	0.6423	0.6433	0.6485	0.6466	0.6402	0.6401	0.6353	0.6309
JAC	0.6177	0.6314	0.6421	0.6416	0.6459	**0.6478**	**0.6433**	**0.6434**	**0.6394**	0.6331
IG	0.6292	0.6317	0.6345	0.6361	0.6370	0.6346	0.6316	0.6340	0.6248	0.6327
CHI	**0.6342**	0.6315	0.6358	0.6348	0.6331	0.6310	0.6137	0.6150	0.6220	0.6161
RMIJ	0.5816	0.5887	**0.6446**	**0.6491**	**0.6510**	0.6445	0.6422	0.6381	0.6379	0.6327
GSS	0.5805	0.6129	0.6229	0.6359	0.6309	0.6262	0.6261	0.6194	0.6166	0.6339
PCC	0.6082	0.6171	0.6340	0.6356	0.6160	0.6101	0.6012	0.6015	0.6038	0.6153
MI	0.2254	0.2259	0.4115	0.3778	0.5326	0.5673	0.5853	0.5712	0.5675	**0.6346**
LOR	0.3431	0.4642	0.5747	0.5903	0.5929	0.5769	0.5739	0.5689	0.5667	0.6166
max	**0.6342**	0.6318	0.6446	0.6491	**0.6510**	0.6478	0.6433	0.6434	0.6394	0.6346

으로 산출한 것이다. 여기서 상대적 상호정보량J(RMIJ)이 50%의 자질집합을 사용하는 경우에 최고 성능(0.6510)이지만, 카이제곱통계량(CHI)은 가장 적은 자질집합(10%)을 사용하는 경우에도 비교적 좋은 성능을 보이는 것으로 나타났다.

〈표 4〉는 9개 자질선정 방법별로 자질집합을 증가하는 경우에 복수-범주 할당 방법에 대한 랜덤포레스트(RF)의 분류 성능을 mac_F1으로 산출한 것이다. 단일-범주 할당 방법에서와 마찬가지로 카이제곱통계량(CHI)이 가장 적은 자질집합(10%)으로 최고 성능이었으며(0.7819), 상대적으로 소규모의 자질집합(10%~40%)을 사용하는 경우에 지속적으로 좋은 성능을 보였다. 이에 반해 문헌빈도(DF), 상대적 상호정보량J(RMIJ), 상호정보량(MI)은 더 많은 자질집합(50% 이상)을 사용하여야 좋은 성능을 보이는 것으로 나타났다.

〈표 5〉는 복수-범주 할당 방법에 대한 랜덤포레스트(RF)의 분류 성능을 mic_F1으로 산출한 것이다. 카이제곱통계량(CHI)이 가장 적은 자질집합(10%)으로 최고 성능이면서(0.7983),

〈표 4〉 자질선정을 적용한 랜덤포레스트(RF) 분류 성능: 복수-범주, mac_F1

구분	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
DF	0.7211	0.7212	0.7214	0.7188	0.7188	**0.7148**	0.7072	**0.7068**	0.6940	0.6856
JAC	0.7208	0.7210	0.7163	0.7164	0.7099	0.7065	0.7068	0.6967	0.6975	0.6908
IG	0.7268	0.7189	0.7141	0.7096	0.7019	0.7047	0.6939	0.6820	0.6800	0.6902
CHI	**0.7819**	**0.7530**	**0.7440**	**0.7342**	0.7059	0.6983	**0.7105**	0.6990	0.6945	0.6877
RMIJ	0.7511	0.7123	0.7215	0.7183	**0.7225**	0.7122	0.7028	0.7067	0.6945	0.6863
GSS	0.6132	0.6706	0.6830	0.6836	0.6876	0.6860	0.6866	0.6807	0.6824	0.6917
PCC	0.6012	0.6054	0.6461	0.6821	0.6762	0.6925	0.6881	0.6758	0.6922	0.6942
MI	0.3189	0.3184	0.3007	0.4510	0.4548	0.5845	0.6275	0.6755	**0.7019**	**0.6961**
LOR	0.3281	0.3397	0.4788	0.5930	0.6780	0.6859	0.6840	0.6720	0.6602	0.6913
max	**0.7819**	0.7530	0.7440	0.7342	0.7225	0.7148	0.7105	0.7068	0.7019	0.6961

〈표 5〉 자질선정을 적용한 랜덤포레스트(RF) 분류 성능: 복수-범주, mic_F1

구분	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
DF	0.7836	0.7821	0.7785	0.7725	0.7725	**0.7671**	0.7534	0.7515	0.7419	0.7361
JAC	0.7835	0.7808	0.7720	0.7682	0.7619	0.7535	0.7544	**0.7535**	**0.7524**	0.7362
IG	0.7884	0.7768	0.7672	0.7630	0.7559	0.7491	0.7387	0.7349	0.7226	0.7366
CHI	**0.7983**	**0.7874**	**0.7830**	**0.7763**	0.7648	0.7455	**0.7579**	0.7502	0.7380	0.7322
RMIJ	0.7775	0.7808	0.7755	0.7756	**0.7814**	0.7647	0.7528	0.7520	0.7416	0.7338
GSS	0.7320	0.7398	0.7519	0.7528	0.7491	0.7412	0.7400	0.7339	0.7323	0.7401
PCC	0.7273	0.7339	0.7510	0.7672	0.7517	0.7503	0.7464	0.7395	0.7433	0.7397
MI	0.4028	0.4035	0.3671	0.5517	0.5494	0.6528	0.6575	0.6734	0.6875	**0.7402**
LOR	0.3440	0.4013	0.5071	0.6571	0.7161	0.7340	0.7203	0.7096	0.7025	0.7371
max	**0.7983**	**0.7874**	**0.7830**	0.7763	**0.7814**	0.7671	0.7579	0.7535	0.7524	0.7402

소규모의 자질집합(10%~30%)을 사용하는 경우에 안정적으로 좋은 성능을 보이는 것으로 나타났다. 이는 자동분류 관련 선행연구에서 랜덤포레스트(RF)가 상대적으로 소규모의 자질집합에서 더 나은 성능을 보이며, 특히 카이제곱통계량(CHI)이 좋은 성능을 보인 경우가 많은 것과 일치한다(Roul & Rai, 2016). 결과적으로, 랜덤포레스트(RF)를 적용한 국내 학술지 논문의 자동분류에서는 카이제곱통계량(CHI)이 소규모의 자질집합으로 높은 성능을 기대할 수 있는 자질선정 방법인 것으로 나타났다.

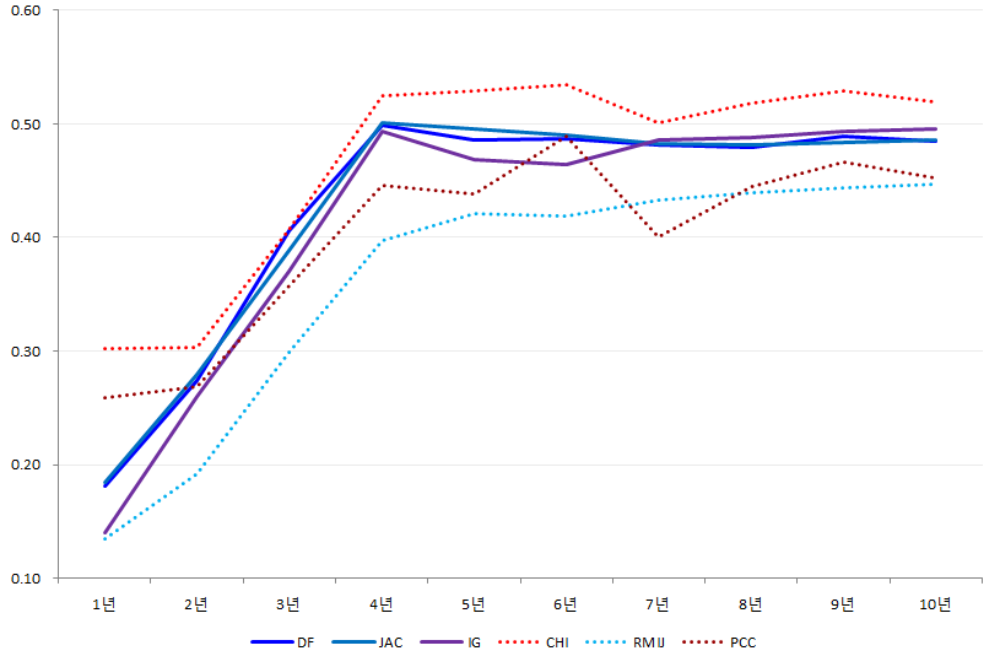
4.3 학습집합 크기

이전 실험에서 랜덤포레스트(RF)에 9개 자질선정 방법을 적용한 대부분의 경우에, 카이제곱통계량(CHI)이 가장 적은 자질집합(10%)으로 최고 성능을 보이는 것으로 나타났다. 이에 따라 이전 실험의 9개 자질선정 방법 중에서 좋은 성능을 보인 상위의 6개 방법으로 선정한 최소의 자질집합(10%)을 사용하면서, 학습집합의 크기를 년차별(1년~10년)로 증가시키는 경우에 두 가지 범주 할당 방법에 따른 성능을 살

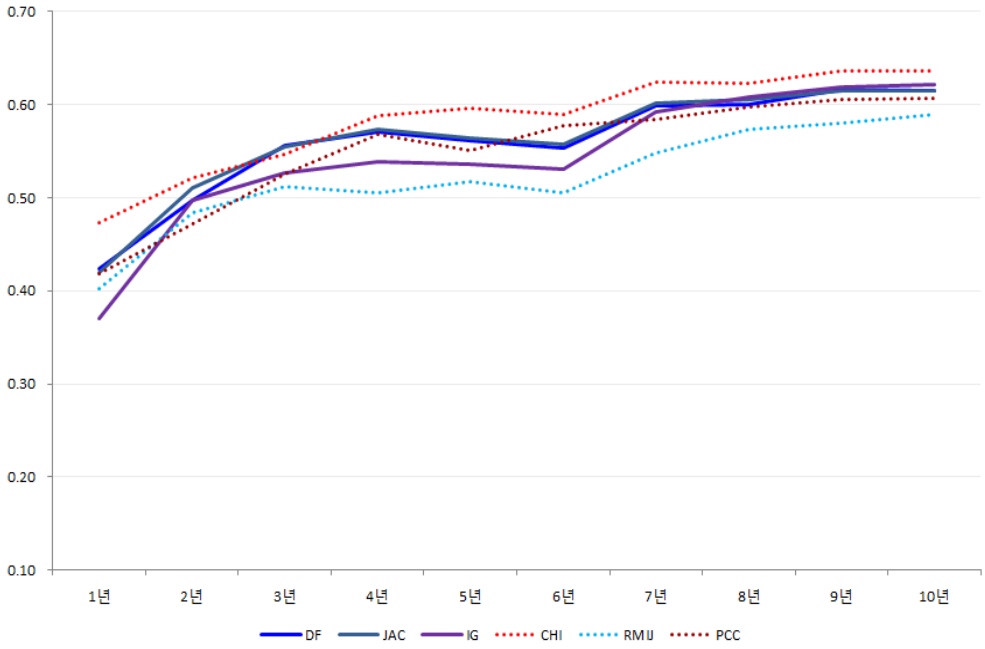
펴보았다.

〈그림 3〉은 6개 자질선정 기법으로 추출한 최소한의 자질집합(10%)을 사용하고 학습집합의 크기를 최근 1년부터 전체 10년까지 증가시키는 가운데 단일-범주를 부여한 분류 성능을 mac_F1으로 산출한 결과이다. 여기서 6개 자질선정 기법 모두가 4년 이상의 학습집합을 사용하였을 때 상당히 좋은 성능을 보였고, 최고 성능은 카이제곱통계량(CHI)으로 선정한 10%의 자질집합과 9년의 학습집합을 사용한 것이었다(0.5339). 또한 동일한 조건에서 단일-범주를 부여한 랜덤포레스트(RF)의 성능을 mic_F1으로 산출한 것은 〈그림 4〉이다. 6개 자질선정 기법 모두 대부분의 학습집합을 사용하는 경우에 가장 좋은 성능이었고, 최고 성능은 카이제곱통계량(CHI)에 기초한 10%의 자질집합과 9년의 학습집합을 사용한 것이었다(0.6359).

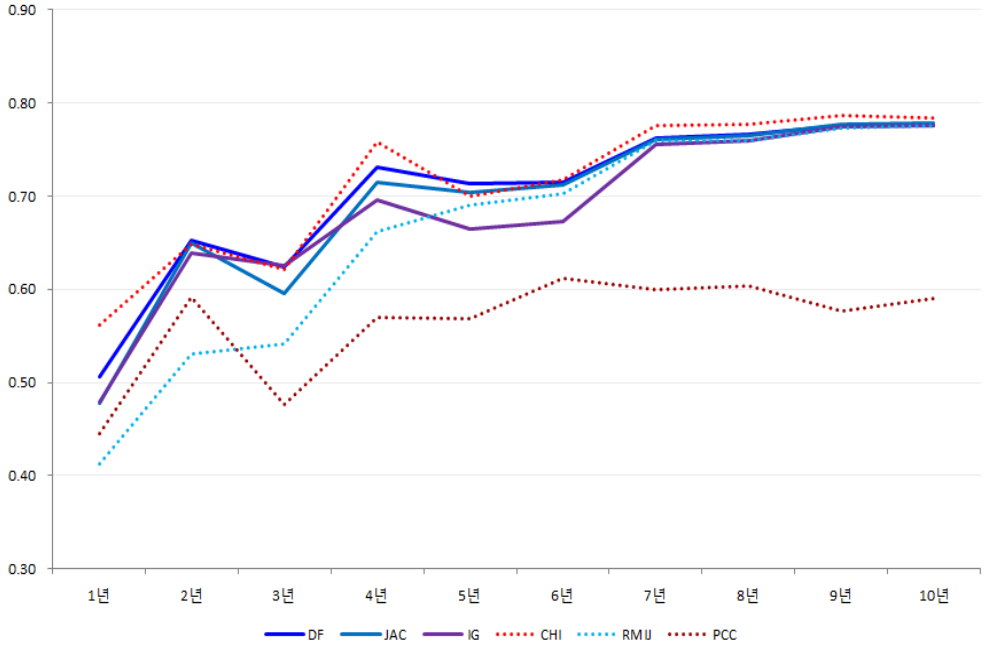
〈그림 5〉는 6개 자질선정 방법으로 선정한 최소한의 자질집합(10%)을 사용하고 학습집합의 크기를 변화시키면서, 복수-범주를 부여한 분류 성능을 mac_F1으로 산출한 결과이다. 대부분의 자질선정 방법이 전체 학습집합(10년)을 사용한 경우에 가장 좋은 성능을 보였고, 최고 성능은 카이제곱통계량(CHI)으로 선정한 10%의 자질집합과 전체 10년의 학습집합을 사용한 것이었다(0.7860). 동일한 조건에서 복수-범주를 부여한 랜덤포레스트의 성능을 mic_F1으로 산출한 〈그림 6〉에서도 대부분이 전체 학습집합(10년)을 사용하였을 때 가장 좋은 성능이었으며, 최고 성능은 카이제곱통계량(CHI)으로 선정한 상위 10%의 자질집합과 전체 10년의 학습집합을 사용한 것이었다(0.7998).



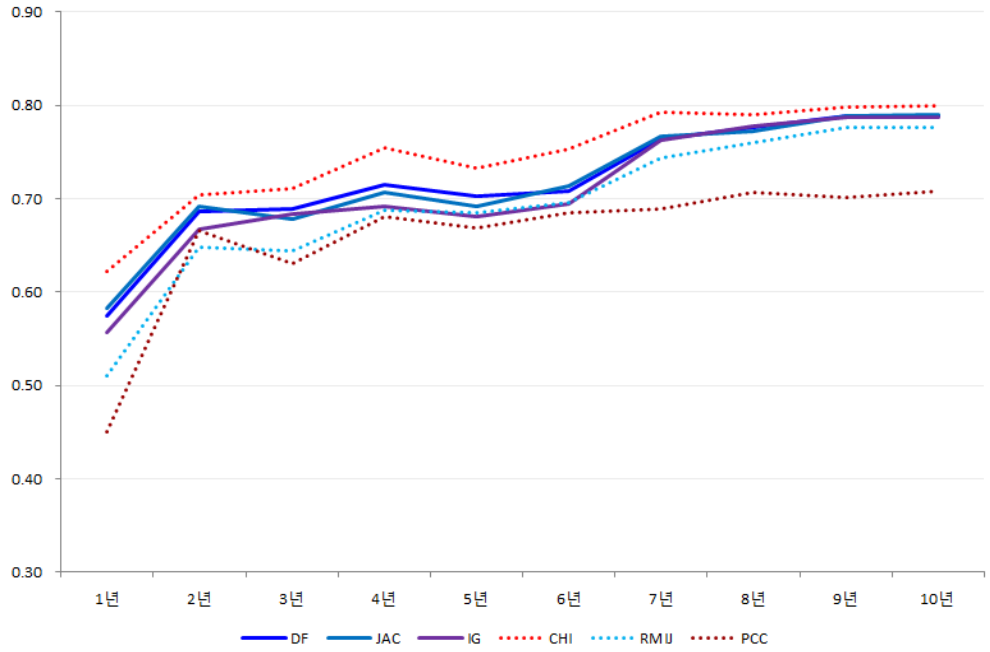
〈그림 3〉 학습집합 크기에 따른 랜덤포레스트(RF) 분류 성능: 단일_범주, mac_F1



〈그림 4〉 학습집합 크기에 따른 랜덤포레스트(RF) 분류 성능: 단일_범주, mic_F1



〈그림 5〉 학습집합 크기에 따른 랜덤포레스트(RF) 분류 성능: 복수_범주, mac_F1



〈그림 6〉 학습집합 크기에 따른 랜덤포레스트(RF) 분류 성능: 복수_범주, mic_F1

4.4 종합 분석

본 연구에서 설정한 연구 문제를 중심으로 실험 결과를 종합적으로 분석한 결과는 다음과 같다. 첫째, 국내 학술지 논문의 자동분류에서 트리 수의 변화에 따라 랜덤포레스트(RF)의 성능에 차이가 있는 지를 분류 성능(effectiveness)과 컴퓨터 처리의 효율성(efficiency) 측면에서 살펴보았다. 선행연구에서 알려진 바와 같이 분류 성능(mac_F1, mic_F1) 측면에서 랜덤포레스트(RF)는 트리 수가 증가할수록 성능이 향상되었다. 그러나 트리 수 구간을 1~10(A)에서 1000~10000(D)까지 증가시키는 경우에 100~1000(C) 구간에서 가장 크게 성능이 향상되었고, 그 이상으로 트리 수를 늘린 1000~10000(D) 구간과의 성능 차이는 크지 않았다. 또한 처리 시간(ms) 측면에서는 이전의 트리 수 구간에 비해 1000~10000(D) 구간의 평균학습시간이 크게 증가하여 거의 10배 이상의 시간이 소요되었다. 따라서 국내 학술지 논문의 자동분류 목적으로 랜덤포레스트(RF)를 적용하는 경우에 분류 성능과 컴퓨터 처리의 효율성 측면에서 가장 적절한 트리 수는 100~1000(C) 구간이라 할 수 있다.

둘째, 국내 학술지 논문의 자동분류에서 자질선정 기법과 자질집합의 변화에 따라 랜덤포레스트(RF)의 성능에 차이가 있는 가를 분류 성능(mac_F1, mic_F1) 측면에서 살펴보았다. 트리 수 구간을 100~1000(C)으로 설정하고 선행연구에서 많이 사용된 자질선정 방법 9개에 기초하여 자질집합의 비율(10%~100%)을 변화시킨 랜덤포레스트(RF) 분류 실험에서, 전반

적으로 카이제곱통계량(CHI)이 가장 소규모의 자질집합(10%)만으로 최고 성능을 보였다. 따라서 국내 학술지 논문의 자동분류 목적으로 랜덤포레스트(RF)를 이용하는 경우에 분류 성능 측면에서 가장 뛰어난 자질선정 방법과 자질집합의 비율은 카이제곱통계량(CHI)에 기초한 소규모의 자질집합(10%)이라 할 수 있다.

셋째, 국내 학술지 논문의 자동분류에서 학습집합의 크기에 따라 랜덤포레스트(RF)의 성능에 차이가 있는 가를 분류 성능(mac_F1, mic_F1)을 중심으로 살펴보았다. 트리 수 구간을 100~1000(C)으로 고정하고 이전 실험에서 좋은 성능을 보인 상위 6개 방법으로 선정한 최소한의 자질집합(10%)을 사용하면서 학습집합의 크기를 년차별(1년~10년)로 증가시킨 랜덤포레스트(RF)의 성능을 살펴보았다. 여기서 6개 자질선정 방법 모두 대부분의 학습집합(9년~10년)을 사용하였을 때 좋은 성능을 보였고, 특히 카이제곱통계량(CHI)으로 선정한 소규모의 자질집합(10%)과 9년 이상의 학습집합을 사용한 경우가 최고 성능이었다. 결과적으로, 국내 학술지 논문의 자동분류에 랜덤포레스트(RF)를 적용할 때는 트리 수 구간 100~1000(C), 카이제곱통계량(CHI)으로 선정한 소규모의 자질집합(10%), 그리고 대부분의 학습집합(9년~10년)을 사용하는 경우에 가장 좋은 분류 성능을 기대할 수 있다.

5. 결 론

학술정보의 생산 및 유통의 폭발적 증가로 인해 디지털 큐레이션의 한 분야로서 컴퓨터에 의한 자동분류 또는 자동주석(automatic classification or automatic annotation)의 필요성이 나날이 증가하고 있다. 이에 따라, 전문가에 의한 수작업 분류를 지원할 수 있는 효과적인 방법으로 기계학습 알고리즘에 기초한 자동분류를 적극적으로 모색할 필요가 있다. 실제 환경에서 학술지 논문의 분류는 하나의 논문에 다수의 범주가 할당되는 복수-범주 분류에 해당하며, 그 결과로 할당된 범주정보는 범주별 문헌 분포가 균등하지 않은 불균형 데이터(imbalanced data)인 경우가 대부분이다. 따라서 본 연구는 실제 환경의 불균형 데이터로서 국내 학술지 논문의 자동분류에 대표적인 앙상블 기법인 랜덤포레스트(RF)를 적용하는 방안을 모색하였다.

실제 환경의 불균형 데이터인 국내 학술지 논문집합에 두 가지 범주 할당 방법(단일-범주, 복수-범주)으로 범주를 자동 할당하는 환경에서, 랜덤포레스트(RF)의 주요 성능 요소인 트리 수, 자질선정, 학습집합 크기에 기초한 연구 문제를 설정하고 각 요소에 대한 다각적인 실험을 수행하였다. 이러한 세 가지 연구문제를 중심으로 실험 결과를 종합적으로 분석한 결과, 국내 학술지 논문의 자동분류에 랜덤포레스트(RF)를 적용할 때는 트리 수 구간 100~1000(C), 카이제곱통계량(CHI)으로 선정한 최소한의 자질집합(10%), 그리고 대부분의 학습집합(9년~10년)을 사용하는 경우에 가장 좋은 분류 성능을 기대할 수 있는 것으로 나타났다.

본 연구의 결과는 특정 분야의 학술지에 수록된 논문을 대상으로 실험한 것이므로 전체 학문분야로 일반화하기에는 어려움이 있다. 따라서 다른 학술지 또는 학문분야로 실험 문헌집합을 확장할 필요가 있다. 또한, 텍스트 범주화(문헌

의 자동분류)에 좋은 성능을 보이는 것으로 알려진 다른 분류 알고리즘과의 비교를 위한 추가적인 연구도 필요할 것이다.

참 고 문 헌

강수연, 전희정, 김지혜, 송정우 (2015). 국내 드라마 시청률 예측 및 영향요인 분석. 응용통계연구, 28(5), 933-949. http://dx.doi.org/10.5351/KJAS.2015.28.5.933

권안나 (2013). 랜덤포레스트를 이용한 변수 선택. 석사학위논문, 인하대학교 대학원, 통계학과.

김성진, 안현철 (2016). 기업신용등급 예측을 위한 랜덤 포레스트의 응용. 산업혁신연구, 32(1), 187-211.

김판준 (2006). 기계학습을 통한 디스크립터 자동부여에 관한 연구. 정보관리학회지, 23(1), 279-299. https://doi.org/10.3743/KOSIM.2006.23.1.279

김판준 (2016). 기계학습에 기초한 자동분류의 성능 요소에 관한 연구. 정보관리학회지, 33(2), 33-59. http://dx.doi.org/10.3743/KOSIM.2016.33.2.033

김판준 (2018). 기계학습에 기초한 국내 학술지 논문의 자동분류에 관한 연구. 정보관리학회지, 35(2), 37-62. https://doi.org/10.3743/KOSIM.2018.35.2.037

남승현, 오명섭, 김성관, 강창완, 김규곤, 최승배 (2017). 사용자 성향 그룹 분류를 위한 머신러닝 모델 비교, Journal of the Korean Data Analysis Society, 19(5), 2501-1507.

서종덕 (2016). 데이터 마이닝 기법을 이용한 환율예측 - GARCH와 결합된 랜덤 포레스트 모형. 산업경제연구, 29(5), 1607-1628.

유진은 (2015). 랜덤 포레스트: 의사결정나무의 대안으로서의 데이터 마이닝 기법. 교육평가연구, 28(2), 427-448.

윤태균, 이관수 (2008). 의료진단 및 중요 검사 항목 결정 지원 시스템을 위한 랜덤 포레스트 알고리즘 적용. 전기학회논문지, 57(6), 1058-1062.

이재윤 (2005). 자질 선정 기준과 가중치 할당 방식간의 관계를 고려한 문서 자동분류의 개선에 대한 연구. 한국문헌정보학회지, 39(2), 123-146. http://dx.doi.org/10.4275/kslis.2005.39.2.123

이진욱, 유국현, 문병민, 배석주 (2017). 감성분석과 Word2vec을 이용한 비정형 품질 데이터 분석. 품질경영학회지, 45(1), 117-127. http://dx.doi.org/10.7469/JKSQM.2017.45.1.117

이현주, 신동규, 박희원, 김수한, 신동일 (2011). 부정맥 증상을 자동으로 판별하는 Random Forest 분류기의 정확도 향상을 위한 수정 알고리즘에 대한 연구. 정보처리학회논문지B, 18(6), 341-348.

정석원, 최맹식, 김학수 (2016). 랜덤 포레스트를 이용한 한국어 상호참조 해결. 정보처리학회논문지: 소프트웨어 및 데이터 공학, 5(11), 535-540.

정준호, 장경현, 김재협 (2016). 랜덤포레스트와 유전알고리즘을 이용한 표적 분류 기법. 2016년 대한전자공학회 추계학술대회 논문집, 601-604.

조현채, 박철용 (2018). 랜덤포레스트를 이용한 신문사들의 19대 대통령 선거 보도 특성 분석. 한국데이터정보과학회지, 29(2), 367-375. http://dx.doi.org/10.7465/jkdi.2018.29.2.367

최혁진, 최성욱, 한경숙 (2012). Random forest를 이용한 단백질에서의 DNA 결합 부위 예측. 정보과학회논문지: 소프트웨어 및 응용, 39(7), 515-522.

홍준혁, 고병철, 남재열 (2013). 가중치 기반 Bag-of-Feature와 앙상블 결정 트리를 이용한 정지 영상에서의 인간 행동 인식. 한국통신학회논문지, 38(1), 1-9. https://doi.org/10.7840/kics.2013.38A.1.1

Afianto, M. F., Adiwijaya, & Al-Faraby, S. (2017). Text categorization on Hadith Sahih Al-Bukhari using Random Forest, International Conference on Data and Information Science, IOP Conference Series: Journal of Physics: Conf. Series 971. http://doi.org/10.1088/1742-6596/971/1/012037

Amaratunga, D., Cabrera, J., & Lee, Y. (2008). Enriched random forests. Bioinformatics, 24(18), 2010-2014. https://doi.org/10.1093/bioinformatics/btn356

Aung, W. T., Myanmar, Y., & Hla, K. H. M. S. (2009). Random forest classifier for multi-category classification of web pages. In Services Computing Conference, APSCC 2009. IEEE Asia-Pacific, 372-376. http://doi.org/10.1109/APSCC.2009.5394100

Austin, P. C., Tu, J. V., Ho, J. E., Levy, D., & Lee, D. S. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: A case study examining classification of heart failure subtypes. Journal of Clinical Epidemiology, 66(4), 398-407. http://doi.org/10.1016/j.jclinepi.2012.11.008

Berk, R., Li, A., & Hickman, L. J. (2005). Statistical difficulties in determining the role of race in capital cases: A re-analysis of data from the state of Maryland. Journal of Quantitative Criminology, 21(4), 365-390. https://doi.org/10.1007/s10940-005-7354-7

Boinee, P., Angelis, A. D., & Foresti, G. L. (2005). Meta random forests. International Journal of Computational Intelligence, 2(3), 138-147.

Brandenburg, Minke (2017). Text classification of Dutch police records. Unpublished master's thesis, Utrecht University Artificial Intelligence, Netherlands.

Breiman L. (2002). Random forests. Machine Learning, 45(1), 5-32.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446-3453. http://doi.org/10.1016/j.eswa.2011.09.033

Choi, S., & Kim, H. (2016). Tree size determination for classification ensemble. Journal of the Korean Data &Information Science Society, 27(1), 255-264. http://dx.doi.org/10.7465/jkdi.2016.27.1.255

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792. http://doi.org/10.1890/07-0539.1

Dogan, T., & Uysal, A. K. (2018). The impact of feature selection on urban land cover classification. International Journal of Intelligent Systems and Applications in Engineering(IJISAE), 6(1), 59-64. http://doi.org/10.18201/ijisae.2018637933

Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: From early developments to recent advancements. Systems Science & Control Engineering, 2(1), 602-609. http://doi.org/10.1080/21642583.2014.956265

Gao, D., Zhang, Y., & Zhao, Y. (2009). Random forest algorithm for classification of multiwavelength data. Research in Astronomy and Astrophysics, 9(2), 220-226. http://doi.org/101088/1674-4527/9/2/011

Kim, M. J., Kang, D. K., &. Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. Expert Systems with Applications, 42(3), 1074-1082. https://doi.org/10.1016/j.eswa.2014.08.025

Klassen, M., & Paturi, N. (2010). Web document classification by keywords using Random Forests. In: Zavoral F., Yaghob J., Pichappan P., El-Qawasmeh E. (eds) Networked Digital Technologies. NDT 2010. Communications in Computer and Information Science, vol 88. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14306-9_26

Kong, Q., Gong, H., Ding, X., & Hou, R. (2017). Classification application based on mutual information and Random Forest method for high dimensional data. 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, 171-174. https://doi.org/10.1109/IHMSC.2017.45

Latinne, P., Debeir, O., & Decaestecker, C. (2001). Limiting the number of trees in Random Forests. In: Kittler J., Roli F. (eds) Multiple Classifier Systems. MCS 2001. Lecture Notes in Computer Science, vol 2096. Springer, Berlin, Heidelberg, 178-187. https://doi.org/10.1007/3-540-48219-9_18

Lee, Jaesung, & Kim, Dae-Won (2015). Mutual information-based multi-label feature selection using interaction information. Expert Systems with Applications, 42(4), 2013-2025. https://doi.org/10.1016/j.eswa.2014.09.063

Liparas D., HaCohen-Kerner Y., Moumtzidou A., Vrochidis S., & Kompatsiaris I. (2014). News articles classification using Random Forests and weighted multimodal features. In: Lamas D., Buitelaar P. (eds) Multidisciplinary Information Retrieval. IRFC 2014. Lecture Notes in Computer Science, vol 8849. Springer, Cham. https://doi.org/10.1007/978-3-319-12979-2_6

Lok, C. (2010). Speed reading. Nature 463, 28. http://doi.org/10.1038/463416a

Löw, F., Schorcht, G., Michel, U., Dech, S., & Conrad, C. (2012). Per-field crop classification in irrigated agricultural regions in middle Asia using random forest and support vector machine ensemble. Proc. SPIE 8538, Earth Resources and Environmental Remote Sensing/GIS Applications III, 85380R (25 October 2012). http://doi.org/10.1117/12.974588

Ma, L. (2017). A multi-label text classification framework: Using supervised and unsupervised feature selection strategy. Unpublished doctoral dissertation, Georgia State University. retrieved from https://scholarworks.gsu.edu/cs_diss/134

Ma, L., & Fan, S. (2017). CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. BMC Bioinformatics, 18(1), 169. https://doi.org/10.1186/s12859-017-1578-z

Ma, L., Zhang, Y., Sunderraman, R., Fox, P., Laird, A., Turner, J., & Turner, M. (2015). Hybrid feature selection methods for online biomedical publication classification. 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, Canada, 1-8. https://doi.org/10.1109/CIBCB.2015.7300320

Madjarov, G., Kocev, D., Gjorgjevikj, D., & Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. Pattern Recognition, 45, 3084-3104. https://doi.org/10.1016/j.patcog.2012.03.004

Manning, Christopher, Raghavan, & Prabhakar (2008). Introduction to information retrieval. NY, USA: Cambridge University Press.

Nayak, S., Ramesh, R., & Shah, S. (2013). A study of multi-label text classification and the effect of label hierarchy. CS224N Project Report, USA: Stanford University. retrieved from https://nlp.stanford.edu/courses/cs224n/2013/reports/nayak.pdf

Robnik-Šikonja M. (2004). Improving Random Forests. In: Boulicaut JF., Esposito F., Giannotti F., Pedreschi D. (eds) Machine Learning: ECML 2004. ECML 2004. Lecture Notes in Computer Science, vol 3201. Springer, Berlin. https://doi.org/10.1007/978-3-540-30115-8_34

Roul, R. K., & Rai, P. (2016). A new feature selection technique combined with elm feature

space for text classification. In Proceedings of the 13th International Conference on Natural Language Processing, 285-292.

Siroky, D. S. (2009). Navigating random forests and related advances in algorithmic modeling. Statistics Surveys, 3, 147-163.

Trieschnigg, D., Pezik, P., Lee, V., Jong, F. D., Kraaij, W., & Rebholz-Schuhmann, D. (2009). MeSH Up: Effective MeSH text classification for improved document retrieval. Bioinformatics, 25(11), 1412-1418. https://doi.org/10.1093/bioinformatics/btp249

Tsymbal, A., Pechenizkiy, M., & Cunningham, P. (2006) Dynamic integration with random forests. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds) Machine Learning: ECML 2006. ECML 2006. Lecture Notes in Computer Science, vol 4212. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11871842_82

Turner, M. D., Chakrabarti, C., Jones, T. B., Xu, J. F., Fox, P. T., Luger, G. F., Laird, A. R., & Turner, J. A. (2013). Automated annotation of functional imaging experiments via multi-label classification. Frontiers in neuroscience, 7, 240. http://doi.org/10.3389/fnins.2013.00240

Ward, M. S., Pajevic, J., Dreyfuss, J., & Malley, J. (2006). Short-term prediction of mortality in patients with systemic lupus erythematosus: Classification of outcomes using random forests. Arthritis and Rheumatism, 55(1), 74-80. http:/doi.org/10.1002/art.21695

Wu, Q., Ye, Y., Zhang, H., Ng, M. K., & Ho, Shen-Shyang. (2014). Fores texter: An efficient random forest algorithm for imbalanced text categorization. Knowledge-Based System, 67, 105-116. http://doi.org/10.1016/j.knosys.2014.06.004

Xu, B., Guo, X., Ye, Y., & Cheng, J. (2012). An improved random forest classifier for text categorization. Journal of Computers, 7(12), 2913-2920. http://dx.doi.org/10.4304/jcp.7.12.2913-2920.

Xu, B., Huang, J. Z., Williams, G., & Ye, Y. (2012). Hybrid weighted random forests for classifying very high dimensional data. International Journal of Data Warehousing and Mining, 8(2), 44-63. http://dx.doi.org/10.4018/jdwm.2012040103

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In Proceedings of the Fourteenth International Conference on Machine Learning, July 08-12, 412-420.

Yao, D., Yang, J., & Zhan, X. (2013). An improved random forest algorithm for class-imbalanced data classification and its application in PAD risk factors analysis. The Open Electrical & Electronic Engineering Journal, 7, (Supple 1: M7) 62-70.

http://dx.doi.org/10.2174/1874129001307010062

Zhou Q., Zhou H., & Li, T. (2016). Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features. Knowledge-Based Systems, 95, 1-11. https://doi.org/10.1016/j.knosys.2015.11.010

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Choi, H., Choi, S., & Han, K. (2012). Prediction of DNA binding sites in proteins using a Random Forest. Journal of KIISE, 39(7), 515-522.

Hong, J., Ko, B., & Nam, J. (2013). Human action recognition in still image using weighted bag-of-features and ensemble decision trees. The Journal of Korean Institute of Communications and Information Sciences, 38(1), 1-9. https://doi.org/10.7840/kics.2013.38A.1.1

Jeong, J., Jang, K., & Kim, J. (2016). Target classification method using Random Forest and genetic algorithm. 2016 IEIE Fall Conference, 601-604.

Jeong, S., Choi, M., & Kim, H. (2016). Coreference resolution for Korean using Random Forests. Journal of KIISE, 5(11), 535-540.

Jo, H., & Park, C. (2018). Analysis of reporting characteristics of newspapers in the 19th presidential election based on random forest. Journal of the Korean data & information science society, 29(2), 367-375. http://dx.doi.org/10.7465/jkdi.2018.29.2.367

Kang, S., Jeon, H., Kim, J., & Song, J. (2015). A study on domestic drama rating prediction. The Korean Journal of Applied Statistics, 28(5), 933-949. http://dx.doi.org/10.5351/KJAS.2015.28.5.933

Kim, P. J. (2006). A Study on automatic assignment of descriptors using machine learning. Journal of the Korean Society for information Management, 23(1), 279-299. https://doi.org/10.3743/KOSIM.2006.23.1.279

Kim, Pan Jun (2016). An analytical study on performance factors of automatic classification based on machine learning. Journal of the Korean Society for information Management, 33(2), 33-59. http://dx.doi.org/10.3743/KOSIM.2016.33.2.033

Kim, Pan Jun (2018). An analytical study on automatic classification of domestic journal articles based on machine learning. Journal of the Korean Society for information Management, 35(2), 37-62. https://doi.org/10.3743/KOSIM.2018.35.2.037

Kim, S., & Ahn, H. (2016). Application of Random Forests to corporate credit rating prediction.

The Journal of Business and Economics, 32(1), 187-211.

Kwon, A. (2013). Variable selection using Random Forest. unpublished master's thesis, Inha University.

Lee, C., Yoo, K., Mun, B., & Bae, S. (2017). Informal quality data analysis via sentimental analysis and Word2vec method. Journal of Korean Society for Quality Management, 45(1), 117-127. http://dx.doi.org/10.7469/JKSQM.2017.45.1.117

Lee, H., Shin, D., Park, H., Kim, S., & Shin, D. (2011). Research on the modified algorithm for improving accuracy of Random Forest classifier which identifies automatically arrhythmia. The KIPS Transactions: Part B, 18(6), 341-348.

Lee, Jae-Yun (2005). An empirical study on improving the performance of text categorization considering the relationships between feature selection criteria and weighting methods. Journal of the Korean Society for Library and Information Science, 39(2), 123-146. http://dx.doi.org/10.4275/kslis.2005.39.2.123

Nam, S., Oh, M., Kim, S., Kang, C., Kim, G., & Choi, S. (2017). Comparison of machine learning models for classification into user-oriented groups. Journal of the Korean Data Analysis Society, 19(5), 2501-1507.

Suh, J. (2016). Foreign exchange rate forecasting using the GARCH extended Random Forest model. Journal of Industrial Economics and Business, 29(5), 1607-1628.

Yoo, J. (2015). Random forests, an alternative data mining technique to decision tree. Journal of Educational Evaluation, 28(2), 427-448.

Yun, Taegyun, & Yi, Gwan-Su (2008). Application of Random Forest algorithm for the decision support system of medical diagnosis with the selection of significant clinical test. The Transactions of the Korean Institute of Electrical Engineers, 57(6), 1058-1062.