The Korean Journal of Applied Statistics
2024, Vol. 37, No. 5, 691–702

DOI: https://doi.org/10.5351/KJAS.2024.37.5.691
Print ISSN1225-066X / Online ISSN 2383-5818

Ensemble model through mixed projections useful for big data analytics

Hyejoon Park[a], Hyunjoong Kim[1,a], Yung-Seop Lee[2,b]

[a]Department of Applied Statistics, Yonsei University; [b]Department of Statistics, Dongguk University

Abstract

In this paper, we propose mixed projection forest (MPF), a new classification ensemble method that can be effectively applied in the field of big data analysis. When training individual classifiers within an ensemble, MPF uses oblique hyperplanes using combined rotation matrix derived from data projection techniques of principal component analysis (PCA) and canonical linear discriminant analysis (CLDA), thereby improving the accuracy of each classifier. Additionally, the diversity of individual classifiers is improved by generating various rotation matrices through random partitioning of the input variable set. This approach ultimately enhances classification performance and proves to be highly effective in big data analysis that demands precision. We conducted a performance comparison of MPF with existing classification ensemble models using 30 real or simulated datasets. The results indicate that MPF achieves competitive performance in terms of classification accuracy and classifier diversity.

Keywords: classification, ensemble, rotation forest, canonical forest, random rotation ensemble

1. 서론

분류 앙상블 모형은 빅데이터 분석에서 분류 성능 향상을 위해 널리 사용되고 있다. 이는 여러 개의 약한 분류기를 생성하고 그 결과를 종합하여 최종 예측 결과를 도출하는 방식으로, 하나의 우수한 분류기로 결과를 예측하는 것보다 높은 정확도를 지닌다. 약한 분류기로는 의사결정나무가 가장 많이 선택되고 있다.

대표적인 앙상블 모형으로는 bagging (Breiman, 1996)과 random forest (Breiman, 2001)가 있다. 하지만, 이 모형들은 초평면(hyperplane)을 수직 또는 수평 형태로만 형성하는 한계점이 있다. 이러한 한계를 해결하기 위해 rotation forest (Rodríguez 등, 2006), canonical forest (Chen 등, 2014), 그리고 random rotation ensemble

Hyunjoong Kim's work was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) support program (IITP-2023-00259934) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) and by the National Research Foundation of Korea (NRF) grant funded by the Korean government (No. 2016R1D1A1B02011696). Yung-Seop Lee's work was supported by the National Research Foundation(NRF) grant funded by the Korea government (MSIT) (No.2021R1A2C1007095) and by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-2020-0-01789) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation).

[1]Corresponding author: Department of Applied Statistics, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea. E-mail: hkim@yonsei.ac.kr

[2]Corresponding author: Department of Statistics, Dongguk University, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, Korea. E-mail: yung@dongguk.edu

Published 31 October 2024 / journal homepage: http://kjas.or.kr

(Blaser와 Fryzlewicz, 2016)등의 모형들이 제안되었다. 이들 모형은 붓스트랩 데이터로 분류기를 구축하는 기존 모형들을 변형하여, 회전 행렬에 의해 회전된 데이터로 분류기를 구축함으로써 사선 형태의 초평면을 형성한다. 각 모형은 principal component analysis (PCA) (Jolliffe, 2002), canonical linear discriminant analysis (CLDA) (Fukunaga, 2013) 및 난수로 생성한 투영축을 만드는 과정에서 계산되는 회전 행렬로 데이터를 회전시킨후 앙상블의 분류기들을 학습시킨다.

이 논문에서는 PCA와 CLDA의 조합으로 데이터를 회전하는 새로운 앙상블 모형인 mixed projection forest (MPF)를 제안한다. MPF는 PCA와 CLDA를 통해 계산된 투영 정보를 활용하여 회전 데이터를 생성한 후 개별 분류기를 학습시켜 앙상블 내 각 분류기의 정확도를 향상시키고자 한다. 또한, 다양한 투영 정보를 선형 조합하여 회전 데이터를 생성함으로써 개별 분류기의 다양성도 확보할 수 있다. 본 논문의 2절에서는 MPF 모형이 회전 행렬을 생성하는 방법과 이를 이용하여 회전된 데이터로 개별 분류기를 구축하고 최종 범주를 예측하는 모델링 과정을 설명한다. 이후 비교 실험을 통해 MPF와 기존 앙상블 모형들과의 분류 성능을 비교하고 결론을 도출하고자 한다.

2. Mixed projection forest 알고리즘

분류기의 다양성을 증가시키기 위해, 1절에서 언급된 모형들은 각각 다른 접근법을 채택한다. Bagging은 붓스트랩 데이터로 개별 분류기를 구축하는 방법을 사용한다. Random forest는 이 방법을 확장하여 붓스트랩 데이터를 사용하면서 의사결정나무의 각 노드마다 무작위로 선택된 입력 변수만을 사용하여 자식 노드를 생성하는 방법을 채택한다.

반면, rotation forest, canonical forest, 그리고 random rotation ensemble은 회전 행렬을 학습 데이터에 적용하여 새로운 회전된 데이터를 생성한 후, 이를 기반으로 개별 분류기를 더 다양하게 구축하는 방법을 사용한다. 다만, 회전 행렬을 계산하는 과정에서 각기 다른 방식으로 투영축을 계산한다. 새로운 앙상블 모델인 mixed projection forest (MPF) 역시 데이터를 회전시켜 분류기의 다양성을 높이는 방식을 사용하는데, 기존의 회전 데이터를 사용하는 모델과는 회전 행렬 생성 방법에서 차이가 있다. 2절에서는 MPF의 알고리즘을 구체적으로 설명하고, 기존 모델들과의 차이점을 논의할 것이다.

2.1. MPF의 회전 행렬

Rotation forest는 PCA 변환 행렬을, canonical forest는 CLDA 변환 행렬을, 그리고 random rotation ensemble은 변환 행렬의 각 요소를 난수로 구성하여 이를 회전 행렬로 사용한다. Rotation forest와 canonical forest는 데이터 내 변동성 정보를 반영하여 데이터를 회전시키므로 개별 분류기의 정확도는 높아지지만, 개별 분류기의 다양성은 크게 향상되지 않는 경향이 있다. 반면, random rotation ensemble은 무작위 값을 사용함으로써 분류기의 다양성만을 증가시킨다는 특징이 있다.

MPF는 PCA와 CLDA의 투영 축을 선형 조합하여 회전 행렬을 생성함으로써, 개별 분류기의 정확성을 크게 저하시키지 않으면서도 다양성을 증가시키는 것을 목표로 한다. MPF 방법에서 회전 행렬의 생성 과정은 다음과 같이 요약할 수 있다

1. 우선 학습 데이터의 입력 변수에 대해 PCA를 수행하고, 이어서 입력 변수와 목표 변수 모두를 포함하여 CLDA를 적용한다. 여기서 입력 변수는 p개의 요소로 구성되어 있다.
2. PCA와 CLDA의 변환 행렬은 p개의 투영 축으로 구성되어 있으며, i번째 투영 축은 각각 $\mathbf{pca.v}_i \in \mathbb{R}^p$와 $\mathbf{clda.v}_i \in \mathbb{R}^p$로 표기한다.
3. 0에서 1 사이의 난수 α를 추출한다.

4. p차원 회전 행렬 $\mathbf{R}$은 PCA와 CLDA 투영 축의 선형 조합으로 구성되며, i번째 열은 $\mathbf{v}_\mathrm{i}$로 표기한다.

$$\mathbf{v}_\mathrm{i} = \alpha \times \mathbf{pca.v}_\mathrm{i} + (1-\alpha) \times \mathbf{clda.v}_\mathrm{i},$$
$$\mathbf{v}_\mathrm{i} = \frac{\mathbf{v}_\mathrm{i}}{\|\mathbf{v}_\mathrm{i}\|},$$
$$\boldsymbol{R} = [\mathbf{v}_1, \ldots, \mathbf{v}_\mathrm{p}] \in \mathbb{R}^{p\times p}.$$

PCA는 입력 변수의 공분산을 가장 잘 보존하는 투영 축을 제공하며, CLDA는 클래스 간 분리가 최대가 되는 공간으로 투영되는 회전 행렬을 계산한다. MPF는 이 두 가지 투영 축을 조합하여, 데이터의 전역적 및 국소적 구조를 모두 반영하는 새로운 변환 데이터를 생성함으로써 분류기의 다양성을 증대시키고자 한다. 이를 통해 개별 분류기의 정확도를 크게 저하시키지 않으면서도 다양성을 높여 앙상블 전체의 정확도를 향상시키는 효과를 기대할 수 있다.

2.2. MPF 분류기 생성 알고리즘

앙상블 모형에서는 가급적 다양한 예측을 할 수 있는 여러 개의 개별 분류기를 생성하여야 한다. MPF 방법에서는 2.1절에 언급한 다양한 회전 행렬을 활용하는 방법 외에도, 입력 변수의 부분집합을 사용하여 더욱 다양한 개별 분류기를 생성한다. 이 방법은 rotation forest (Rodríguez 등, 2006)와 canonical forest (Chen 등, 2014)에서도 유사하게 사용되었다.

MPF 앙상블에서 m번째 개별 분류기를 구축하는 과정은 다음과 같다. 여기서, 데이터는 입력 변수 데이터 $\boldsymbol{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n\times p}$ (n: 관측치 수, $\mathbf{x}_i = [x_{1i}, x_{2i}, \ldots, x_{ni}]^T \in \mathbb{R}^n$) 와 C개의 범주 값$(1, \ldots, C)$을 지닌 목표 변수 $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$로 구성된다고 하자.

1. 앙상블 내 m번째 분류기를 생성하기 위해, p개의 입력 변수를 랜덤하게 K개의 부분집합으로 나눈다. 각 입력 변수는 하나의 부분집합에만 속할 수 있으며, 각 부분집합에는 p/K개의 입력 변수가 있다. 만약 p가 K의 배수가 아니라면, 먼저 $\lfloor p/K \rfloor$개의 입력 변수를 모든 부분집합에 동일하게 할당한다. 그리고 나머지 입력 변수들은 첫 번째 부분집합부터 순차적으로 하나씩 할당된다. 본 논문에서는 Rodríguez 등 (2006)과 Chen 등 (2014)에 따라 $K = \lfloor p/3 + 0.5 \rfloor$를 사용한다.
2. j번째 부분집합에 속하는 입력 변수만으로 구성된 하위 데이터셋 $\boldsymbol{X}_j$를 생성한다. 즉, $\boldsymbol{X}$는 K개의 하위 데이터셋으로 분할된다.
3. 각 하위 데이터셋 $\boldsymbol{X}_j$를 $0.75 \times n$의 크기로 붓스트랩하여 $\boldsymbol{X}_j^{'}$를 생성한다.
4. 각 하위 데이터셋 $\boldsymbol{X}_j^{'}$의 PCA 투영 축과 CLDA 투영 축의 선형 조합으로 회전 행렬 $\boldsymbol{R}_{mj} = [\mathbf{r}_{j1}, \ldots, \mathbf{r}_{jd_j}] \in \mathbb{R}^{d_j\times d_j}$ (d_j: j번째 부분집합에 속한 입력 변수 개수, $\mathbf{r}_{jl} \in \mathbb{R}^{d_j}, \forall l \in \{1, \ldots, d_j\}$)를 생성한다.
5. 각 하위 데이터셋으로 도출한 K개의 회전 행렬로부터 블록 대각 행렬 $\boldsymbol{R}_m$을 생성한다.

$$\boldsymbol{R}_m = \begin{bmatrix} \boldsymbol{R}_{m1} & 0 & \cdots \\ 0 & \ddots & \vdots \\ \vdots & \cdots & \boldsymbol{R}_{mK} \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{r}_{11}, \mathbf{r}_{12}, \ldots, \mathbf{r}_{1d_1} & 0 & \cdots \\ 0 & \ddots & \vdots \\ \vdots & \cdots & \mathbf{r}_{K1}, \mathbf{r}_{K2}, \ldots, \mathbf{r}_{Kd_K} \end{bmatrix} \in \mathbb{R}^{p\times p}.$$

Table 1: The description of 30 datasets

Data	Observations	Categories	Variables	Source
ban	1372	2	4	UCI (Bank note authentication)
bcw	683	2	9	Lim 등 (2000)
blo	748	2	4	UCI (Blood transfusion center)
bos	506	3	12	UCI (Boston housing)
col	366	3	9	UCI (Horse Colic)
con	1473	3	9	KEEL (contraceptive)
cyl	540	2	19	UCI (Cylinder bands)
dia	768	2	8	Loh (2009)
ech	131	2	6	UCI (Echocardiogram)
ger	1000	2	7	UCI (German credit)
hea	270	2	13	KEEL (heart)
int	1000	2	9	Kim 등 (2011)
ion	351	2	32	KEEL (ionosphere)
mar	8777	10	3	Loh (2009)
pho	5404	2	5	KEEL (phoneme)
ring	7400	2	20	KEEL (ring)
rng	1000	2	10	R (mlbench: Ringnorm)
sah	462	2	8	KEEL (saheart)
sat	6435	6	36	UCI (StatLog satellite image)
sea	3000	3	7	Terhune (1994)
snr	208	2	60	R (mlbench: Sonar)
spa	4601	2	57	UCI (Spambase)
spe	267	2	44	UCI (SPECTF heart)
tit	2201	2	3	KEEL (titanic)
trn	1000	2	10	R (mlbench: Threenorm)
two	7400	2	20	KEEL (twonorm)
usn	1302	3	26	Vlachos (2010)
veh	846	4	18	KEEL (vehicle)
vow	990	11	13	KEEL (vowel)
wdbc	569	2	30	KEEL (wdbc)

6. 원 데이터의 변수 순서에 따라 $\boldsymbol{R}_m$의 열을 재배열하여 $\boldsymbol{R}_m^a \in \mathbb{R}^{p\times p}$를 생성한다.
7. 회전 행렬 $\boldsymbol{R}_m^a$를 사용하여 데이터 $\boldsymbol{X}$를 회전시킨다 (= $\boldsymbol{X}\mathbf{R}_m^a$).
8. 회전 데이터 ($\boldsymbol{X}\mathbf{R}_m^a$, $\mathbf{y}$)로 분류기 L_m을 구축한다.

위 과정을 $m = 1, \ldots, M$에 대해 반복하여 총 M개의 분류기를 생성한다. 그런 다음 다수결 규칙을 적용하여 최종 예측 결과를 도출한다. 즉, 데이터 관측치 $\mathbf{x}$의 예측 범주는 다음 수식에 의해 결정된다:

$$L(\mathbf{x}) = \underset{y\in\{1,\ldots,C\}}{\operatorname{argmax}} \sum_{m=1}^{M} I\left(L_m(\mathbf{x}\boldsymbol{R}_m^a) = y\right).$$

3. 비교 실험

Bagging, random forest, rotation forest, canonical forest, random rotation ensemble, 그리고 MPF의 분류 성능을 비교하기 위해 Table 1에 있는 30개의 데이터를 사용해 실험을 진행했다. 30개의 데이터는 주로 UCI 데이터

Table 2: Average misclassification rate for test data (%)

Data	Bagging	Random forest	Rotation forest	Canonical forest	Random rotation ensemble	Mixed projection forest
bal	18.8	15.7	6.2	6.0	10.4	8.7
ban	1.7	1.1	2.0	1.9	1.5	1.5
bcw	3.5	2.9	2.7	2.5	2.8	2.5
blo	26.3	24.6	25.1	24.7	25.1	24.4
bos	21.9	20.9	20.9	20.9	21.1	21.1
col	31.2	29.6	30.1	29.6	30.9	29.9
con	52.6	48.6	53.3	53.5	54.0	53.5
cyl	23.8	22.9	23.4	23.7	30.0	23.4
dia	24.3	24.3	24.0	24.1	24.7	23.9
ech	32.0	30.9	31.6	31.6	31.0	31.6
ger	31.0	29.7	30.3	29.9	32.5	30.2
hea	22.2	21.1	19.4	19.8	20.0	19.6
int	47.4	47.3	48.6	48.4	48.7	48.7
ion	7.7	6.5	5.6	5.8	4.9	5.5
mar	44.1	41.9	44.1	44.0	44.1	44.0
pho	9.8	9.6	9.2	9.7	9.9	9.5
ring	5.5	5.0	3.8	3.1	2.8	3.2
rng	10.2	9.1	7.0	6.7	6.0	6.7
sat	9.5	8.7	8.3	8.4	8.8	8.4
sea	26.8	25.2	25.8	25.7	27.8	25.3
snr	20.6	17.6	15.1	16.4	13.5	15.5
spa	5.8	4.9	4.9	4.8	6.4	4.8
spe	18.6	18.9	18.6	19.3	19.9	19.1
tit	21.2	22.1	21.2	21.2	21.2	21.2
trn	14.6	13.8	12.2	13.6	11.8	12.1
two	3.3	3.0	2.5	2.2	2.3	2.3
usn	25.7	25.9	26.2	26.3	29.2	26.1
veh	24.9	24.7	21.4	20.4	23.6	21.6
vow	9.9	4.6	2.4	3.4	2.7	2.4
wdbc	4.5	4.2	3.1	3.3	2.6	3.2

저장소 (Asuncion와 Newman, 2007) 또는 KEEL 데이터 저장소 (Alcalá-Fdez 등, 2011)에서 가져왔다. Table 1은 왼쪽부터 데이터명, 관측치 수, 목표 변수의 범주 수, 입력 변수 수, 그리고 데이터 출처를 나타낸다. 수치형이 아닌 범주형 변수에는 데이터 회전 개념을 적용할 수 없으므로, 범주형 변수를 제거하고 수치형 변수로만 실험을 진행했다. 실험에 사용된 모든 데이터는 분류에 적합한 데이터이며, 범주형 목표 변수를 가지고 있다.

실험에 사용된 30개의 데이터는 모든 변수의 범위를 동일하게 만드는 최소-최대 정규화로 스케일링되었다. 데이터를 무작위로 70%와 30%의 비율로 나누어 전자는 훈련 데이터, 후자는 평가 데이터로 사용하였으며, 유의성 확인을 위해 해당 과정을 50번 반복했다. 각 데이터별 분류 정확도는 50회 평가 데이터 결과를 이용한 쌍체 검정을 통해 비교되었다. 모든 앙상블 모형은 총 100개의 분류기 결과를 다수결 규칙에 따라 종합하여 평가 데이터의 최종 범주를 예측하였다.

R 프로그램의 randomForest 패키지로 비교 실험을 진행했으며, 자식 노드 생성시 고려되는 후보 입력변수 개수는 random forest에서는 입력 변수 개수 (p)의 제곱근을, random forest를 제외한 방법에서는 p를

Table 3: Win-loss table of accuracy

	Bagging	Random forest	Rotation forest	Canonical forest	Random rotation ensemble	Mixed projection forest
Bagging	-	27 (24)	24 (20)	24 (19)	18 (15)	25 (22)
Random forest	3 (1)	-	18 (14)	18 (14)	13 (12)	18 (15)
Rotation forest	6 (4)	12 (7)	-	15 (8)	11.5 (9)	16.5 (9)
Canonical forest	6 (5)	12 (6)	15 (6)	-	9.5 (8)	18.5 (7)
Random rotation ensemble	12 (8)	17 (13)	18.5 (14)	20.5 (16)	-	20.5 (17)
MPF	5 (2)	12 (5)	13.5 (2)	11.5 (4)	9.5 (6)	-

Table 4: Dominance rank table of accuracy

Method	Wins	Losses	Dominance
MPF	70	19	51
Canonical forest	61	32	29
Rotation forest	56	37	19
Random forest	55	56	−1
Random rotation ensemble	50	68	−18
Bagging	20	100	−80

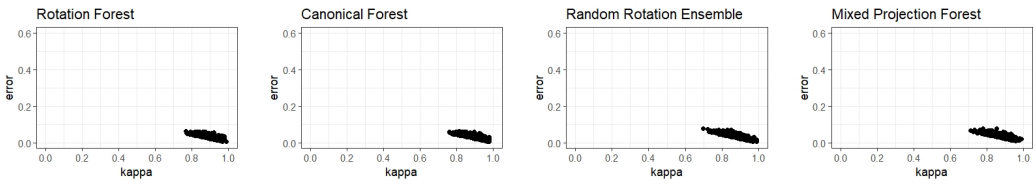
사용했다. Rotation forest, canonical forest 및 MPF에서 입력 변수의 하위 집합 개수 K는 p를 3으로 나누어 반올림한 값으로 설정했다. 회전 행렬을 사용하는 rotation forest, canonical forest, random rotation ensemble 및 MPF는 개별 분류기 생성시 붓스트랩 데이터를 사용하지 않는다. 그 외 매개변수들은 R 패키지의 기본값으로 지정했다.

3.1. 분류 정확도 비교

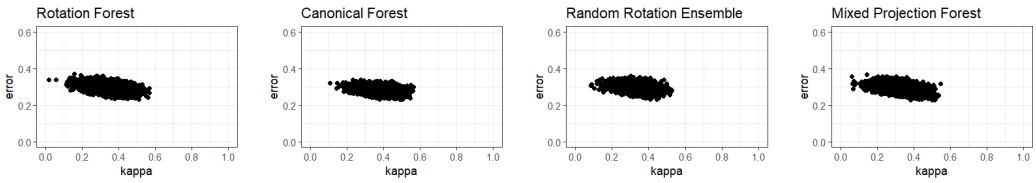
Table 2는 30개 데이터 각각의 오분류율 평균을 정리한 표이다. 오분류율이 낮을수록 분류 성능이 높음을 의미한다. Table 3은 Table 2를 기반으로 30개의 데이터에 대해 여섯 가지 앙상블 모형의 정확도를 비교한 승패표이다. Table 3의 i번째 행의 j번째 열을 a_{ij}라 하자. a_{ij}의 값은 30개의 데이터 중 j번째 열에 적힌 모형이 i번째 행에 적힌 모형보다 분류 성능이 우수한 데이터 개수이다. 괄호 안 값은 쌍체검정을 통해 유의수준 0.05에서 통계적으로 유의하게 분류 성능이 우수한 데이터의 개수를 뜻한다. 예를 들어, a_{16}과 a_{61}의 값은 25 (22)와 5 (2)이다. 이는 MPF와 bagging을 비교했을 때, 30개 데이터 중 25개 데이터에서 MPF의 성능이 높았으며 그중에서 22개에서는 통계적으로 유의하게 높았음을 의미한다. 반대로 30개 데이터 중 5개 데이터에서 bagging의 성능이 높았으며 그중에서 2개에서는 통계적으로 유의하게 높았음을 의미한다. 모형 간에 성능 차이가 없는 경우에는 a_{ij}와 a_{ji} 괄호 밖 값에 0.5를 더한다.

Table 4는 승패 표에서 통계적으로 유의한 결과를 기반으로 구성한 지배 순위(dominence rank)표이다. Table 4의 첫 번째 열은 각 행에 적힌 모형이 다른 모형들에 비해 통계적으로 유의하게 분류 성능이 높은 데이터 수의 합이고, 두 번째 열은 통계적으로 유의하게 분류 성능이 낮은 데이터 수의 합이다. 세 번째 열은 첫 번째 열과 두 번째 열의 차이다. 세 번째 열은 지배 순위를 나타내며 내림차순으로 정렬되어 있다. 표의 상단에 위치할수록 해당 모형은 다른 모형들에 비해 분류 성능이 높음을 의미한다. 따라서, 표의 가장 상단에 위치한 MPF 모형이 다른 다섯 가지 모형에 비해 높은 분류 정확도를 보인다고 해석할 수 있다.

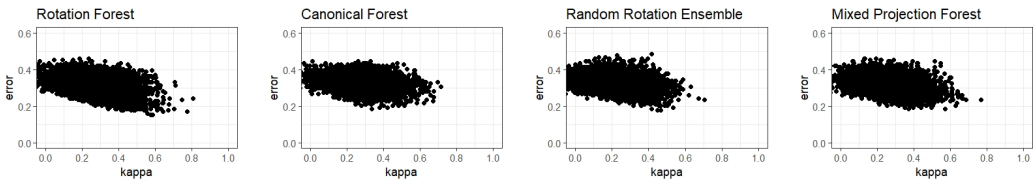
MPF는 rotation forest 및 canonical forest와 비교하여 분류기의 다양성을 향상시키고, random rotation



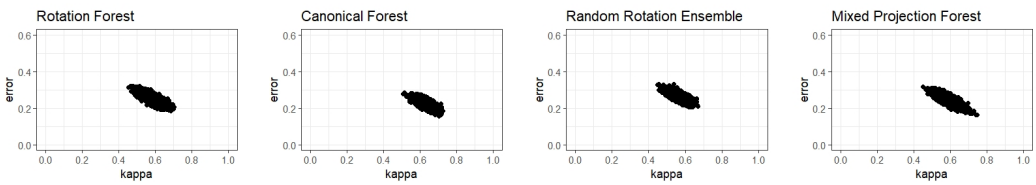
(a) bcw



(b) dia



(c) snr



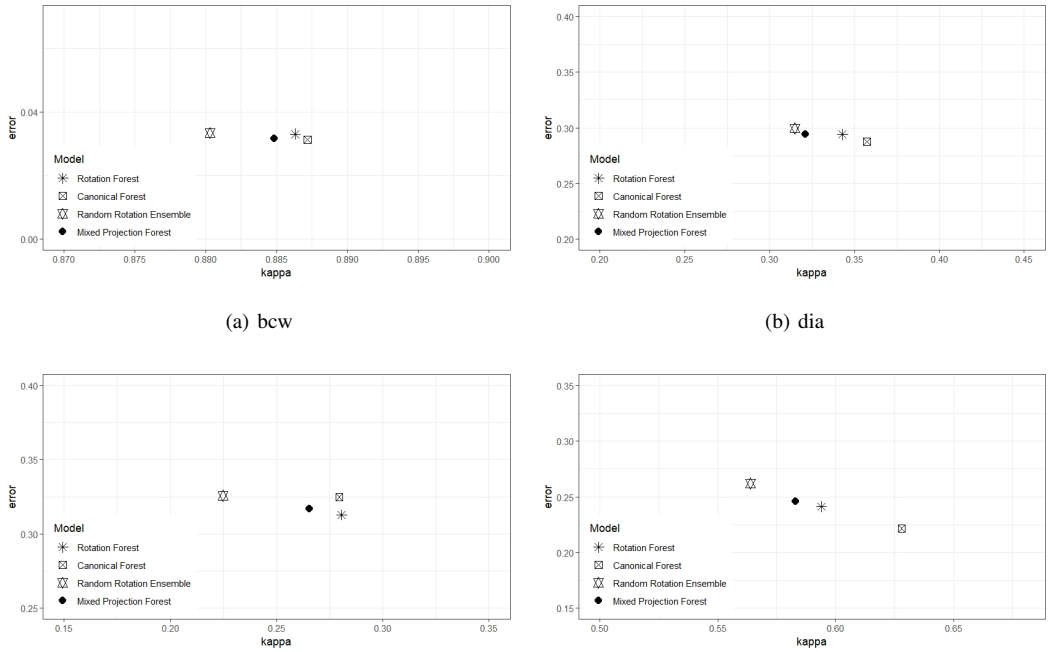
(d) vow

Figure 1: *Kappa-error diagram.*

ensemble과 비교하여 분류기의 정확도를 높임으로써 이들 모델보다 우수한 분류 성능을 나타낸 것으로 판단된다. Random rotation ensemble은 실험에 사용된 모델 중 상대적으로 다양성이 높은 방법임에도 불구하고, 데이터의 고유한 정보를 활용하지 않고 무작위로 데이터를 회전시키기 때문에 분류기의 정확도가 낮아 이러한 결과를 초래한 것으로 생각된다. 다음 장에서 이에 대한 실험결과를 확인하도록 한다.

3.2. 분류기 다양성 비교

앙상블내 분류기들이 얼마나 다양하게 생성되는지 확인하기 위해, 개별 분류기 간의 kappa 통계량 (Cohen, 1960)을 계산해서 이용하고자 한다. Figure 1은 30개의 데이터 중 4개 데이터인 ‘bcw’, ‘dia’, ‘snr’, ‘vow’에 대한 kappa-error 다이어그램이다. 이 그래프의 x축은 kappa 통계량 값들을, y축은 분류기들의 오분류율을 의



(a) bcw (b) dia

(c) snr (d) vow

Figure 2: *Centroid plot of kappa-error diagram.*

미한다. 여기서, kappa 통계량은 M개의 분류기 중에서, 임의의 i번째와 j번째 분류기 L_i와 L_j 사이의 예측 결과 일치도를 의미한다. kappa 통계량 값들이 작을수록 두 분류기의 일치도가 낮아 분류기의 다양성이 높다는 것을 의미한다. 한 데이터에는 총 $\binom{M}{2}$개의 kappa 통계량 값들이 존재한다.

개별 분류기들의 다양성이 높더더라도 오분류율이 높으면 좋은 앙상블 모형이 아니므로, 오분류율도 확인하였다. 오분류율로는 kappa 값을 측정하는 데 사용된 두 분류기 오분류율의 평균을 사용한다. 즉, L_i와 L_j의 오분류율의 평균이 y축 값이 된다. 그래프가 제시되지 않은 26개의 데이터도 Figure 1의 4개 데이터와 유사한 패턴을 나타낸다.

Figure 2는 중심점을 기반으로 Figure 1의 내용을 요약한 것이다. Figure 2의 x축은 kappa 통계량의 평균이고, y축은 오분류율의 평균이다. Figure 2에서 MPF의 kappa 값은 모든 경우에서 rotation forest 또는 canonical forest보다 작다. 즉, MPF는 rotation forest 또는 canonical forest보다 개별 분류기의 다양성이 높다. MPF의 평균 오분류율 값은 rotation forest 및 canonical forest와 유사하거나 상대적으로 높은 편이다. 이는 개별 분류기의 정확도가 다소 낮아질 수 있음을 의미한다. 그러나 3.1절에서 확인한 바와 같이, 앙상블 방법들의 분류 성능 측면에서 MPF는 우수한 결과를 보였다. 이는 개별 분류기의 정확도가 다소 하락하더라도, 보다 다양한 분류기를 생성함으로써 이러한 약점을 충분히 보완했음을 나타낸다.

Table 5와 Table 6는 kappa 통계량 간의 차이를 쌍체 검정을 통해 분석한 결과를 요약한 것이다. Table 6에서 상단에 위치할수록 kappa 통계량이 낮음을 나타내며, 이는 분류기의 다양성이 높음을 의미한다. 회전 행렬을 사용하는 앙상블 모형 중에서 random rotation ensemble은 다른 모형들에 비해 분류기의 다양성이 높으며, MPF, rotation forest, canonical forest 순으로 분류기의 다양성이 높다.

분류기 성능이 분류기의 다양성에 의해 결정된다면, random rotation ensemble의 분류 성능은 Table 4의

Table 5: Win-loss table of Kappa statistic

	Bagging	Random Forest	Rotation Forest	Canonical Forest	Random Rotation Ensemble	Mixed Projection Forest
Bagging	-	29 (29)	11 (10)	12 (12)	14 (14)	14 (14)
Random Forest	1 (1)	-	1 (0)	3 (2)	7 (6)	3 (3)
Rotation Forest	19 (18)	29 (29)	-	14.5 (12)	22.5 (22)	22.5 (22)
Canonical Forest	18 (17)	27 (27)	15.5 (11)	-	21.5 (21)	26.5 (24)
Random Rotation Ensemble	16 (14)	23 (23)	7.5 (5)	8.5 (6)	-	12.5 (11)
MPF	16 (15)	27 (25)	7.5 (5)	3.5 (2)	17.5 (17)	-

Table 6: Dominance rank table of Kappa statistic

Method	Wins	Losses	Dominance
Random Forest	133	12	121
Random Rotation Ensemble	80	59	21
MPF	74	64	10
Bagging	65	79	-14
Rotation Forest	34	100	-66
Canonical Forest	31	103	-72

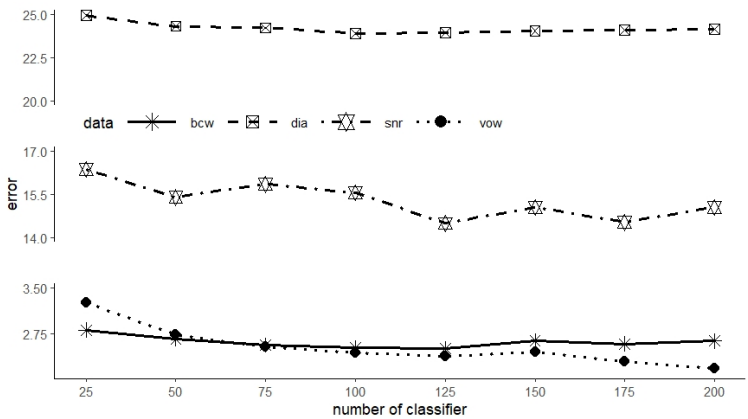


Figure 3: *Changes in misclassification rate by the number of ensembles.*

상단에 위치해야 한다. 그러나 random rotation ensemble은 낮은 분류 성능을 나타냈다. 이로 인해 분류기의 다양성이 분류 성능에 영향을 미치긴 하지만, 이는 절대적인 것은 아니라는 점이 드러난다. 분류 성능을 향상하기 위해서는 분류기의 정확도와 다양성을 균형 있게 높여야 한다. 즉, 분류기의 다양성 효과는 분류기의 정확도가 확보된 경우에 기대할 수 있다.

3.3. 분류기 개수 영향 비교

Figure 3은 Figure 1과 2에 사용된 'bcw', 'dia', 'snr', 'vow' 데이터에 대해, 앙상블에서 생성하는 분류기 개수 M값 변화에 따른 오분류율을 나타낸 것이다. X축은 M값으로 25부터 200까지 25 간격으로 변화시켰고, Y

축은 오분류율의 평균을 나타낸다. Figure 3에서 오분류율은 M값의 상승에 따라 감소하는 경향이 있지만, M 값이 100을 초과하게 되면 그 차이는 미미하다고 볼 수 있다.

4. 결론

이 논문에서는 새로운 분류 앙상블 모델인 mixed projection forest (MPF)를 제안하였다. MPF는 두 가지 관점에서 분류 성능을 향상시킬 수 있었다. 첫 번째는 앙상블 내 개별 분류기의 정확성을 확보하는 것이다. PCA와 CLDA와 같은 데이터 투영 기법의 회전 행렬을 활용하여 생성된 데이터로 개별 분류기를 학습할 때, 사선을 이루는 초평면을 이용한 분류기를 구성하여 앙상블을 형성하였다. 이러한 방식은 직교 초평면을 사용하는 기존의 앙상블 방법보다 더 높은 정확도의 분류기를 생성하는 결과를 가져왔다.

두 번째는 앙상블 내 개별 분류기의 다양성을 증대시키는 것이다. 변수 집합의 랜덤 분할을 통한 입력변수의 부분집합에 PCA와 CLDA에 의한 회전 행렬의 선형 조합을 활용하여 회전 행렬을 도출하는 과정에서, 각 분류기가 생성하는 초평면은 매우 다양하게 구성되었다.

따라서, MPF는 분류기의 다양성과 정확도를 균형 있게 향상시키는 효과적인 분류 앙상블 모델로 결론지을 수 있다. 이러한 결과는 빅데이터 분석에 필요한 분류 문제를 해결하는 데 있어 MPF가 유용한 모델로 활용될 수 있음을 시사한다.

References

Alcalá-FJ, Fernández A, Luengo J, Derrac J, García S, Sánchez L, and Herrera F (2011). KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework, *Journal of Multiple-Valued Logic & Soft Computing*, **17**, 255–287.

Asuncion A and Newman DJ (2007). UCI machine learning repository, Retrieved Oct. 02, 2018, Available from: http://archive.ics.uci.edu/ml/

Blaser R and Fryzlewicz P (2016). Random rotation ensembles, *The Journal of Machine Learning Research*, **17**, 126–151.

Breiman L (1996). Bagging predictors, *Machine Learning*, **24**, 123–140.

Breiman L (2001). Random forests, *Machine Learning*, **45**, 5–32.

Chen YC, Ha H, Kim H, and Ahn H (2014). Canonical forest, *Computational Statistics*, **29**, 849–867.

Cohen J (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, **20**, 37–46.

Fukunaga K (2013). *Introduction to Statistical Pattern Recognition*, Elsevier, Amsterdam, Netherlands.

Jolliffe IT (2002). *Principal Component Analysis for Special Types of Data*, Springer, New York.

Kim H, Kim H, Moon H, and Ahn H (2011). A weight-adjusted voting algorithm for ensembles of classifiers, *Journal of the Korean Statistical Society*, **40**, 437–449.

Lim TS, Loh WY, and Shih Y (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, *Machine Learning*, **40**, 203–228.

Loh WY (2009). Improving the precision of classification trees, *The Annals of Applied Statistics*, **3**, 1710–1737.

Rodriguez JJ, Kuncheva LI, and Alonso CJ (2006). Rotation forest: A new classifier ensemble method, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, 1619–1630.

Terhune JM (1994). Geographical variation of harp seal underwater vocalizations, *Canadian Journal of Zoology*, **72**, 892–897.

Vlachos R (2010). StatLib datasets archive, Retrieved Oct. 02, 2018, Available from: http://lib.stat.cmu.edu/datasets

Received July 30, 2024; Revised August 16, 2024; Accepted August 19, 2024

투영 조합을 통한 빅데이터 앙상블 모형

박혜준[a], 김현중[1,a], 이영섭[2,b]

[a]연세대학교 응용통계학과; [b]동국대학교 통계학과

요 약

이 논문에서는 빅데이터 분석 분야에서 유용하게 사용할 수 있는 새로운 분류 앙상블 방법인 mixed projection forest (MPF)를 제안하였다. 앙상블 내 개별 분류기를 학습할 때, MPF는 주성분 분석(PCA)과 정준 선형 판별 분석(CDA) 등의 데이터 투영 기법의 조합에 의한 회전 행렬을 활용한다. 이를 통해 경사 초평면을 사용함으로써 각 분류기의 정확성을 향상시킨다. 또한 변수 집합의 랜덤 분할을 이용해 다양한 회전 행렬을 도출하여 개별 분류기들의 다양성을 증대시킨다. 이러한 접근 방식은 궁극적으로 분류 성능을 향상시켜 정밀도가 필요한 빅데이터 분석에 매우 효과적이다. 이 논문에서는 실제 및 가상의 30개 데이터셋을 사용하여 MPF와 전통적인 분류 앙상블 모형의 성능을 비교하였다. 결과적으로, MPF는 분류 성능 및 분류기의 다양성 측면에서 우수한 경쟁력을 가진다는 것을 확인할 수 있었다.

주요용어: 분류, 앙상블, rotation forest, canonical forest, random rotation ensemble

[1]교신저자: (03722) 서울특별시 서대문구 연세로 50, 연세대학교 응용통계학과. E-mail: hkim@yonsei.ac.kr
[2]교신저자: (04620) 서울특별시 중구 필동로 1길 30, 동국대학교 통계학과. E-mail: yung@dongguk.edu