PCA, Clustering, Prediction, Visualization

Validity Evaluation and Modeling for Colorimetric Sensor Array

Purpose

- 개발된 비색센서 어레이의 **소화기암 관련 VOC 검출 가능 여부 검증**
- 주어진 자료를 통해 **예측 모형 구축**
- 분석 결과와 새로운 자료에 대한 분석 및 예측 결과 **시각화 툴 개발**

Methodology

- PCA 후 k-means, **계층적 군집분석**
- LDA, Logistic Regression, SVM, XGBoost 모형 이용
- R Shiny 이용한 시각화 웹페이지 개발 후 Docker로 배포

Data

- 총 90개의 VOC를 어레이에 각각 5번 반복하여 노출시켰을 때, 어레이의 35개 spot의 색 변화 데이터
- **VOC 종류(1-90), spot 번호(1-35), 색(R,G,B), 반복 회차(1-5)**

$r_{ijk}, g_{ijk}, b_{ijk}$

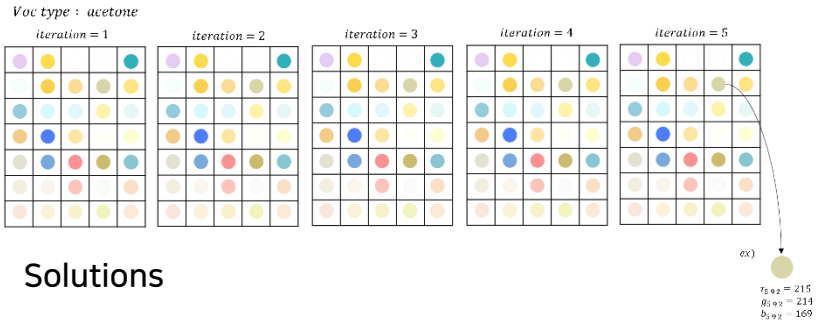
$i = 1,2 \dots, 5\ (number\ of\ iteration)$

$j = 1,2 \dots, 35\ (number\ of\ spot)$

$k = 1,2 \dots, 90\ (VOC\ type)$

Environment

```
R 3.6.1
shiny; shinydashboard; shinymanager; plotly; dendextend
RColorBrewer; MASS; xgboost; e1071
```



Issues

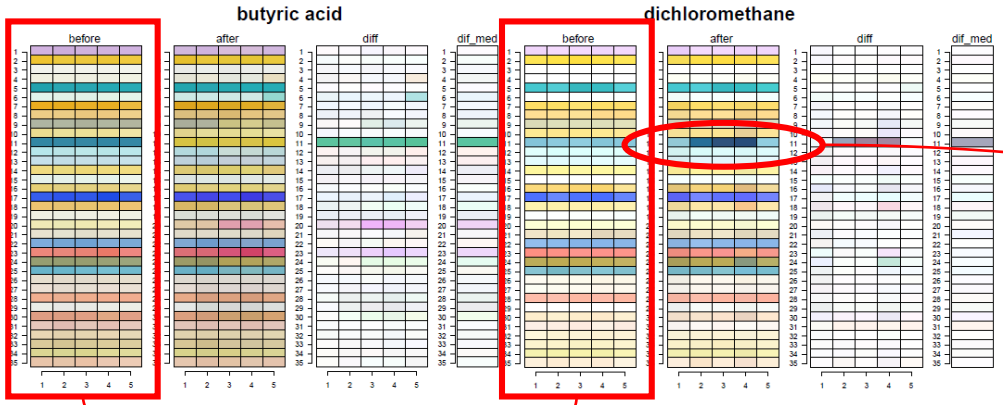
1) VOC 노출과 무관한 **체계적인 편향**에 의한 RGB 값의 변화 관측 (특정 spot에서 큰 변동 관측, 밝기의 정도가 동일하지 않음)
2) 원 데이터는 **실제 환자의 날숨이 아닌** VOC를 각각 노출 시켰을 때의 데이터
3) 예측 모형의 목적에 따라 성능 평가 지표의 중요도 차이 존재

Solutions

1) 문제가 있는 spot의 이미지를 읽어와 전체 spot에서의 RGB 값의 평균값으로 보정(**품질 관리**), 밝기를 보정한 RGB 값으로 변환(**RGB 정규화**)
2) 암 환자와 일반인의 호기가스에서 나오는 **VOC의 상대 비율을 이용해** 암 환자와 일반인 데이터 생성
3) 진단의 목적인 경우, 양성, 음성이라고 예측한 사람들 중 실제로 양성, 음성인 사람의 비율이 중요하다고 판단(**PPV, NPV**)

Validity Evaluation and Modeling for Colorimetric Sensor Array

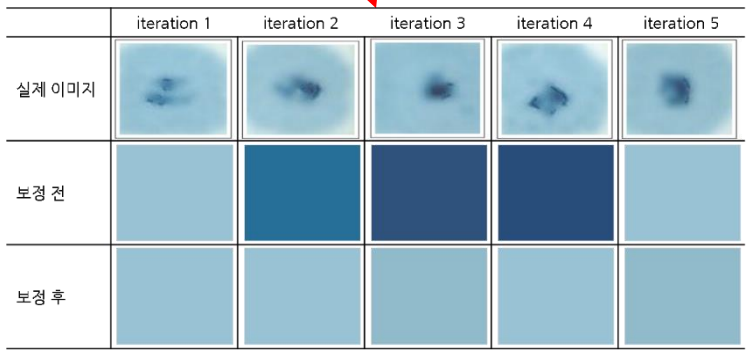
1) 소화기암 관련 VOC 검출 가능 여부 검증

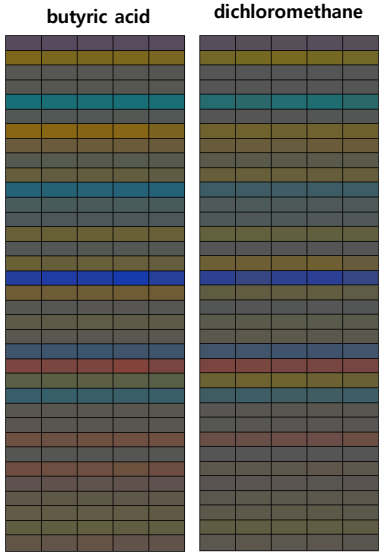


X 축 : $number\ of\ iteration\ (i = 1,2\ldots,5)$
$before$: 5번의 반복실험에서의 노출 전
$after$: 5번의 반복실험에서의 노출 후

Y축 : $number\ of\ spot\ (j = 1,2,\ldots,35)$
$diff$: 5번의 반복실험에서의 노출 전/ 후의 차이
dif_med : 5번의 반복실험에서의 노출 전/ 후 중앙값의 차이

반복 관측치 마다 변동이 큰 spot 보정



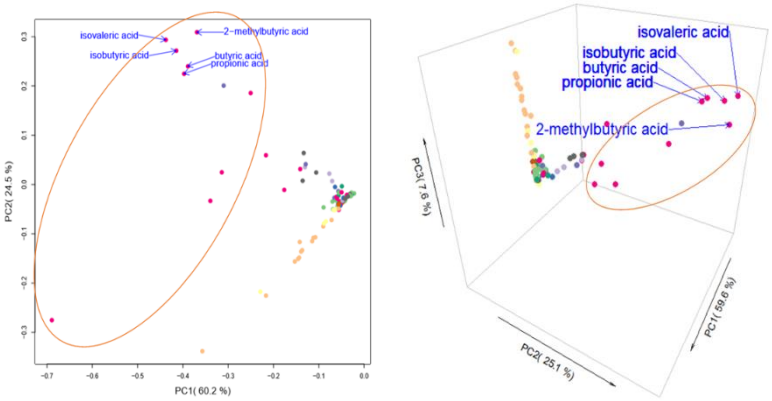


밝기를 보정하기 위해 RGB 정규화

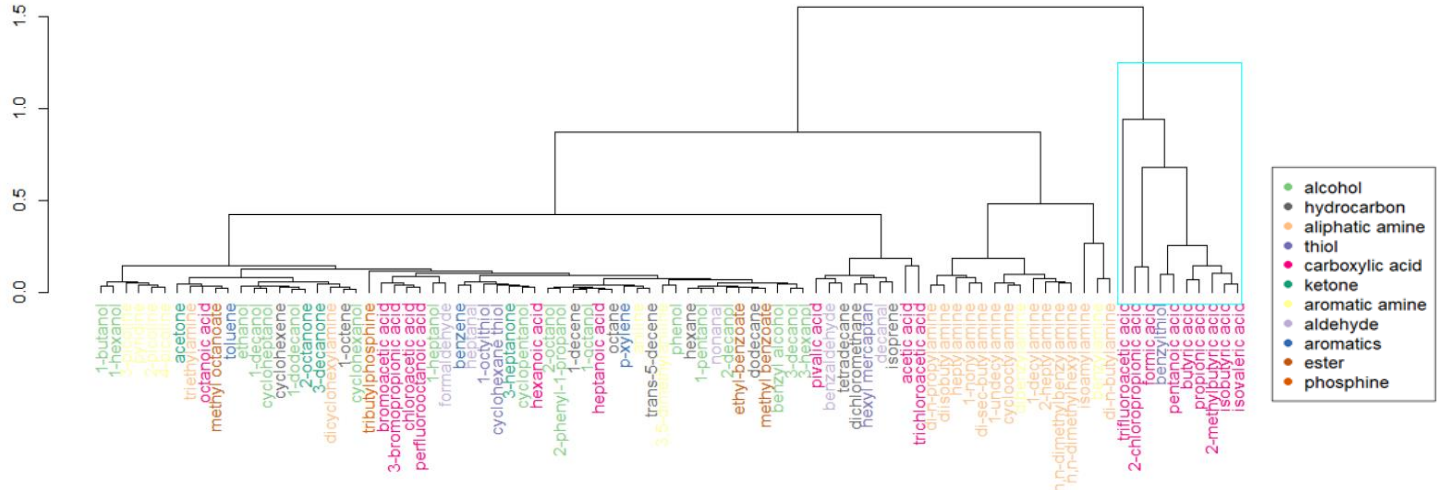
$$\left(\mathbf{r}' = \frac{\mathbf{r}}{\mathbf{r+g+b}},\ \mathbf{g}' = \frac{\mathbf{g}}{\mathbf{r+g+b}},\ \mathbf{b}' = \frac{\mathbf{b}}{\mathbf{r+g+b}}\right)$$

Validity Evaluation and Modeling for Colorimetric Sensor Array

1) 소화기암 관련 VOC 검출 가능 여부 검증



주성분 분석 후 2개의 주성분과 3개의 주성분을 사용했을 때의 산점도



세개의 주성분을 이용하여 계층적 군집분석을 수행한 결과

소화기 암과 관련된 VOC가 하나의 군집으로 묶이는 것을 볼 수 있다.

Validity Evaluation and Modeling for Colorimetric Sensor Array

2) 임의 데이터 생성 후 모델링

① 자료 생성

$r^*_{jk} = \tilde{r}_{.jk} + normal\ (0\ ,\ \hat{\sigma}_{r\ jk\ c}\ )$

$g^*_{jk} = \tilde{g}_{.jk} + normal\ (0\ ,\ \hat{\sigma}_{g\ jk\ c})$

$b^*_{jk} = \tilde{b}_{.jk} + normal\ (0\ ,\ \hat{\sigma}_{b\ jk\ c})$

$j = 1, \dots, 35\ (number\ of\ spot\ ),\ k = 1, \dots, 10\ (voc\ type)\ ,\ c = 1,2,3\ (cluster)$

$normal\ (\mu, \sigma)$: 평균이 μ, 표준편차가 σ 인 정규분포에서 난수 생성

② 호기 가스 상대 비율과 선형 결합

암환자와 일반인을 구분하기 위해 사람의 호기 가스에서 나오는 VOC별 선형 결합

$\underline{w_n} = (w_{1n}, w_{2n}, \dots, w_{10n})'$: 일반인의 호기 가스 상대 비율 $n =$ 일반인

$\underline{w_c} = (w_{1c}, w_{2c}, \dots, w_{10c})'$: 암환자의 호기 가스 상대 비율 $c =$ 암환자

voc.name	w_n	w_c
Isoprene	0.7414	0.7414
Acetone	1	1
Dichloromethane	0.0138	0.0138
ocatane	0.069	0.069
Hexane	0.069	0.069
toluene	0.0862	0.0862
Propionic acid	0.0017	0.4828
Butyric acid	0	0.0345
Isovaleric acid	0.0017	0.0016
2-Methylbutyric acid	0	0.0138

$r^*_{j1} * w_{1n} + r^*_{j2} * w_{2n} + \cdots + r^*_{j10} * w_{10n} \equiv R_{jn}$

$r^*_{j1} * w_{1c} + r^*_{j2} * w_{2c} + \cdots + r^*_{j10} * w_{10c} \equiv R_{jc}$

③ Euclidian distance로 변경

$\underline{d_n} = (d_{1n}, d_{2n}, \dots, d_{35n})'$

$\underline{d_c} = (d_{1c}, d_{2c}, \dots, d_{35c})'$

$d_{jn} = \sqrt{{R_{jn}}^2 + {G_{jn}}^2 + {B_{jn}}^2}\ + normal\ (0\ , \xi\ )$

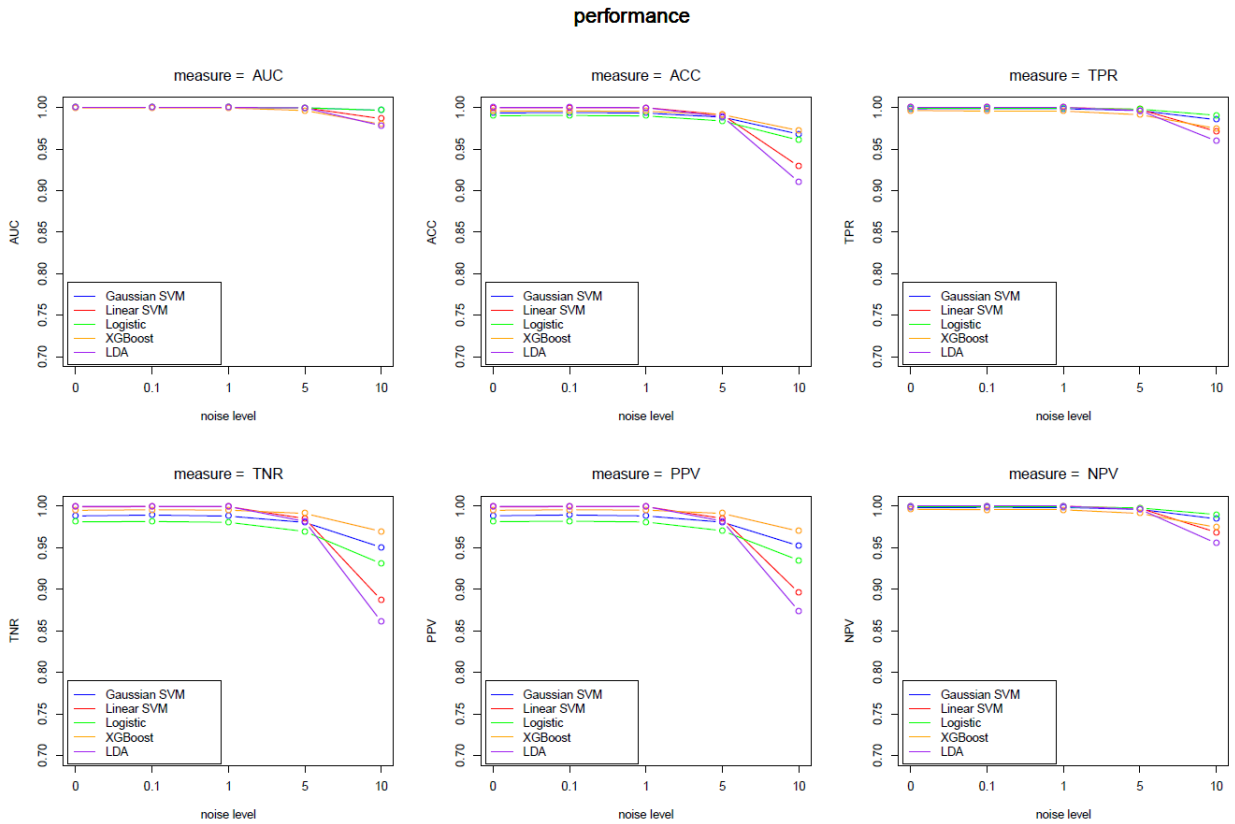
$d_{jc} = \sqrt{{R_{jc}}^2 + {G_{jc}}^2 + {B_{jc}}^2}\ + normal\ (0\ , \xi)$

$j = 1, \dots, 35\ (number\ of\ spot\ ),\ n =$ 일반인 , $c =$ 암환자

$\xi = noise\ \ level\ (0, 0.1, 1, 5, 10)$

Validity Evaluation and Modeling for Colorimetric Sensor Array

2) 임의 데이터 생성 후 모델링



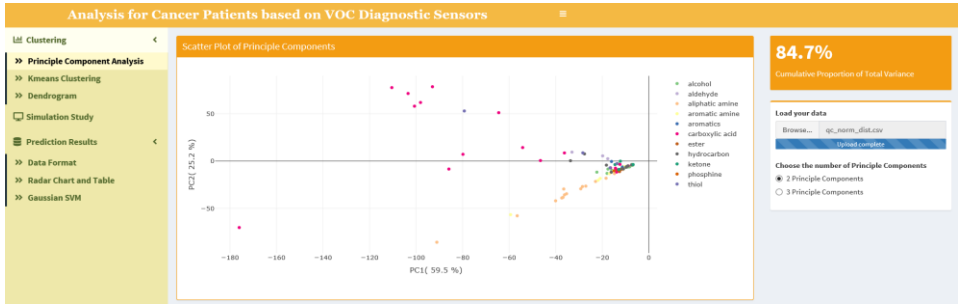
모델링 결과

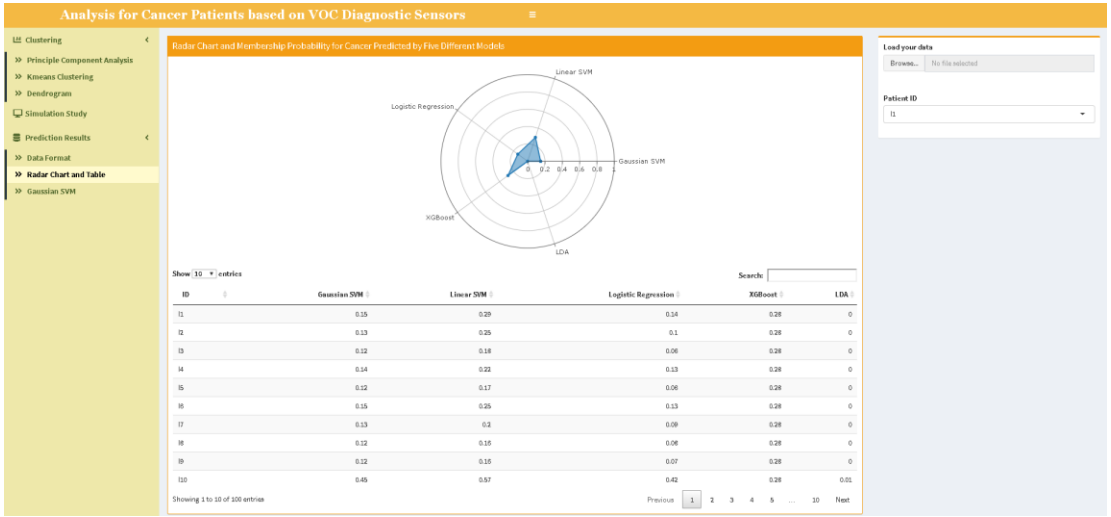
임의로 생성한 자료를 사용했기 때문에 전반적으로 모든 모형이 좋은 성능을 보였다.

PCA, Clustering, Prediction, Visualization

Validity Evaluation and Modeling for Colorimetric Sensor Array

3) 시각화 툴 개발





시각화 툴 화면

연구 결과를 포함하여, 새로운 실험 자료를 업로드하면 해당 자료에 대해주성분 분석, 군집분석을 수행한 결과를 볼 수 있게 구현하였고, 환자의 호기가스를 어레이에 노출시킨 결과를 업로드하면 앞서 구축한 모델로 소화기암 여부를 판단한 결과를 확인할 수 있다.