SVM 이용 제거율 예측시스템

<SVM의 특징>

\* 장점 : 범주나 수치 예측 문제에 대해 사용할 수 있다.

노이즈 데이터에 영향을 크게 받지 않고 잘 과적합화되지 않는다.

특히 잘 지원되는 일부 SVM 알고리즘 때문에 신경망보다 사용하기 쉽다.

높은 정확도와 높은 프로필로 데이터 마이닝 경쟁에서 우승해 인기를 얻었다.

\* 단점 : 최적의 모델을 찾기 위해 커널과 모델에서 매개변수의 여러 가지 조합 테스트가 필요.

특히 입력 데이터 셋이 예제 개수와 속성의 수가 많다면 훈련이 느릴 수 있다.

해석하기 불가능하지 않지만, 어렵고 복잡한 블랙박스를 만든다.

\* SVM은 이미지 데이터 문제를 해결하는 데 매우 적합하다. 노이즈에 민감하지 않게 복잡한 패턴을 학습할 수 있으며, 높은 정확도로 시각 패턴을 인식할 수도 있다. 뿐만 아니라 SVM의 주요 약점에 대해 이미지 처리는 덜 비판적이다.

\* SVM 학습기가 처리하는 모든 속성은 수치여야 하고, 각 속성은 작은 간격 범위로 돼 있어야 한다.

\* 보통 Machine Learning 모형의 경우 종속변수가 클래스를 나타내는 질적 자료이고, 독립변수는 이 클래스의 특징을 나타내는 변수로 양적 변수와 질적 변수를 포함하고 있는 경우에 독립변수를 이용하여 종속변수의 클래스를 예측하는 방법론이다.

<프로젝트에 적용>

입력 데이터(화상검사로부터 나온 데이터) : 디캡 전후 잔여 OCA 좌표 수, 디캡 전후 잔여 OCA 밀도 수준, 디캡 전후 OCA의 무게

총 80개 데이터 중 training data=64개, test data=16개로 8:2의 비율로 사용

1, 잔여 oca라는 전체 데이터 셋 생성, 그 안의 속성으로 oca의 좌표 수, oca의 무게, oca 밀도 수준 수치적으로 나타냄 (이 부분은 java로 화상검사 데이터를 끌고 옴)

2, 잔여 oca 데이터 셋을 8:2의 비율로 training data와 test data로 나눈다.

samp <- c(sample(1:40,32), sample(41:80,32))

3, e1071 함수 사용할 경우 다음과 같이 training set과 test set을 정해줌

oca.tr <- oca[samp,] oca.te <- oca[-samp,]

(이 때 우리가 구하고자 하는 속성 예측 값, 그것이 좌표 수라면 들어가는 training과 test 데이터의 두번째 화상검사 후 데이터도 그 좌표 수에 해당하는 것만 들어감)

4. 각 세부 항목 별 SVM 모델 형성, model1 <- svm(좌표 수~., data=oca 전체 데이터)

model2 <- svm(무게~., data= oca 전체 데이터)

model3 < svm(밀도 수준~., data= oca 전체 데이터)

5, SVM 모델을 실행시켜보면 각 항목(변수)들에 대해서 SVM 수행되는 것 확인 가능

6, 만든 모델과 test 데이터 셋으로 predict 수행

pred <- predict(model1, oca.te, “probabilities”)

7, 예측 한 결과를 table(pred, oca전체데이터$좌표수) 형태로 표로 나타내어 확인 가능