최종 보고서

2019226303 양지혁

1. Introduction

Wisconsin Diagnostic Breast Cancer (WDBC)의 양성 유방 종괴(단순 낭종, 섬유 선종 등)와 악성 유방 종괴(유방암) 두 가지 클래스를 가지는 유방 종괴(breast mass) 데이터는 32개의 데이터 attribute를 가지고 있다. 본 과제에서는 이러한 32개의 attribute를 이용하여 2개의 classification 기법과 1개의 클러스터링 기법을 학습하고 구현 한 후, 테스트 하는 것을 해당 과제의 목표로 한다.

2. Method

해당 데이터를 classification (or clustering)하기 위해서 사용한 method는 다음과 같다.

- supervised ML method : random forest, svm
- unsupervised ML method : k-means

3 가지 머신러닝 기법을 구현하기 위해 python의 scikit library를 사용하였다.

3. Results & Discussion

3.1 Random Forest

3.1.1 Hyperparameter Optimization

- 최적화한 Hyperparameter
  - n_estimators (tree의 수)
  - max_features (최대 선택 할 수 있는 feature)
  - max_depth (최대 tree depth)

  grid-search기법을 기반으로 하여 10-fold cross validation를 수행했을 때, 평균 accuracy가 가장 높은 hyperparameter 조합을 출력하였다. 두 hyperparameter에 대해 조합한 각 hyperparameter 값 리스트는 다음과 같다.
  - ❏ n_estimators = [5, 10, 15, 20, 25, 30, 35, 40, 45, 50]
  - ❏ max_features = [5, 10, 15, 20, 25, 30]
  - ❏ max_depth = [10, 20, 30, 40, 50, 60, 70, 80]

- 결과
  - n_estimators = 50
  - max_features = 10
  - max_depth = 40

※ *3.1.2-3.1.3*의 실험은 해당 결과 파라미터 값들로 설정하여 실험하였음.

3.1.2 classification accuracy

10-fold cross validation를 수행하고, 결과 accuracy list를 평균내어 classification accuracy를 측정하였다. 그 결과 Random forest는 **95.79%**의 accuracy를 보여주었다.

3.1.3 **Discussion** - Most informative feature and information strength

트레이닝 셋 크기는 전체 데이터의 70%, 테스트 셋 크기는 전체 데이터의 30%로 두고, 각 feature의 feature importance를 구했을 때 결과는 그림 1과 같다. 그리고 이 중에서 가장 정보성이 높은 5 가지의 feature를 Table 1로 나타내었다.

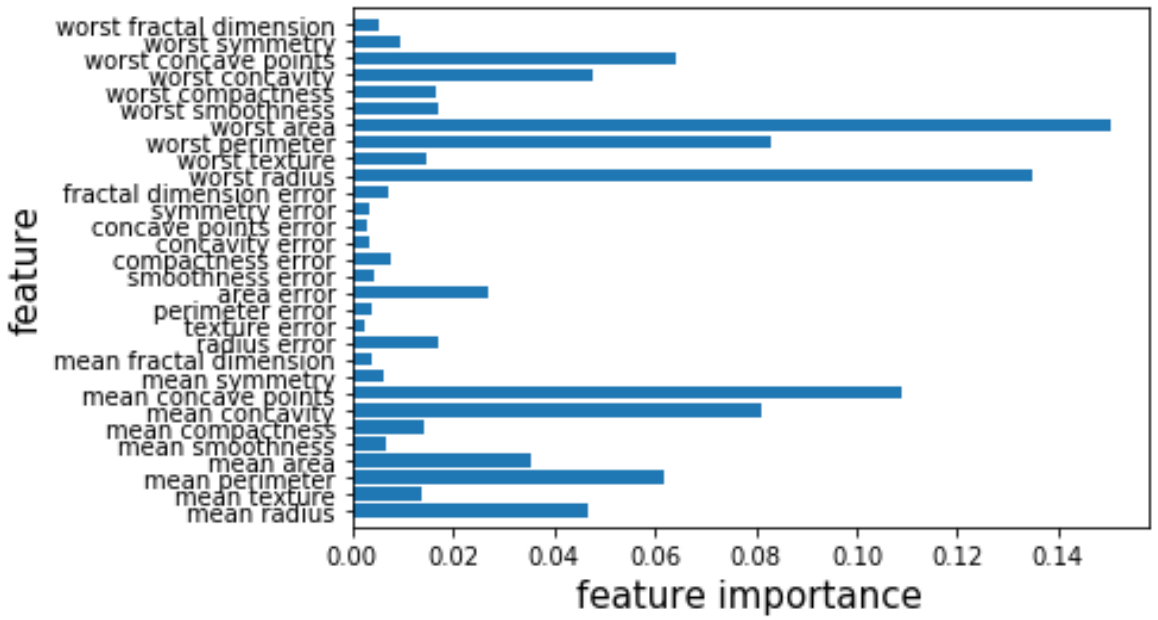

그림 1. Wisconsin Diagnostic Breast Cancer (WDBC)에서 feature 별 importance

Rank	Feature Name (세포 정보)	Information Strength
1	worst_area (가장 크기 큰 세포 3개 면적)	0.150
2	worst_texture (가장 크기 큰 세포 3개 질감)	0.135
3	mean concave points (평균 오목 점 수)	0.109
4	worst perimeter (가장 크기 큰 세포 3개 평균 둘레)	0.083
5	mean concavity (평균 오목도 (오목함의 정도))	0.081

Table 1. 5 features are the most informative

3.2 SVM

3.2.1 Hyperparameter Optimization

SVM 분류기를 만들기 위해 svm의 대표 hyperparameter인 C(cost)를 조정한다. C 값은 커질수록 svm의 hyperplane의 마진 값이 작아지고, 작을수록 hyperplane의 마진 값이 커지는 특징이 있다. 실험은 이전과 똑같이 grid search와 10-fold cross validation을 이용하였다. 대입한 C값의 리스트는 [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]이고, 평균 accuracy가 가장 높은 hyperparameter 값을 출력하였다. 평균 accuracy가 가장 높은 hyperparameter 값은 60이다.

※ *3.2.2-3.2.3*의 실험은 *C*값을 *60*으로 설정하여 실험하였음.

3.2.2 classification accuracy

10-fold cross validation를 수행하고, 결과 accuracy list를 평균내어 classification accuracy를 측정하였다. 그 결과 SVM은 **95.96%**의 accuracy를 보여주었다.

3.2.3 Discussion - accuracy change with different kernel shape

SVM 분류기에서 hyperplane의 커널 모양을 선형 함수가 아닌 Gaussian 함수, sigmoid 함수로 변경 하고 동일한 조건으로 실험 하였을 때, 각각의 분류기의 적중률이 linear일 때보다 높은지 분석해보았다. 그 결과는 Table 2 와 같다. 두 결과값 다 linear 함수를 사용하였을 때 보다 현저히 낮은 accuracy를 가졌다. 이는 Wisconsin Diagnostic Breast Cancer (WDBC) 진단 데이터가 t-SNE로 그림 2와 같이 표현이 되는데, 다른 고차원 함수보다 linear 함수가 overfitting 없이 hyperplane을 설정 할 수 있기 때문이라고 예측된다.

	accuracy
gaussian function	0.6267
sigmoid function	0.6276

Table 2. kernel 함수 변경 시 accuracy

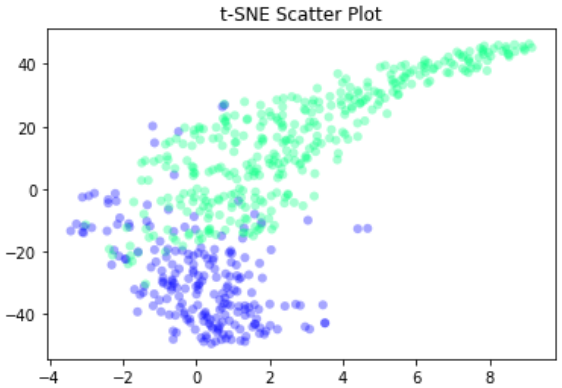

그림 2. WDBC 진단 데이터에 대한 t-SNE 결과 값

3.3 K-means Clustering

3.3.1 rand index, silhouette scores

실험에 사용한 Wisconsin Diagnostic Breast Cancer (WDBC)는 원래 두 개의 label을 가지기 때문에, kmeans를 실행 할 때 cluster 개수를 2개로 두고 실행하여 rand index와 silhouette score값을 구하였다. 그 값은 다음과 같다.

- rand index = 0.4914
- silhouette score = 0.6973

rand index는 매개변수로 WDBC 데이터의 라벨 리스트를 받고, 1에 가까울 수록 WDBC 라벨과 같다는 것을 나타낸다. 그리고 silhouette score는 같은 군집의 데이터 간의 거리와 다른 군집의 데이터 간의 거리를 이용하고, 1에 가까울 수록 잘 클러스터링 된 것을 뜻한다. 실루엣 계수가 0.7, rand index는 약 0.5 정도의 값을 가지는 것으로 클러스터링이 잘 되었다고 말할 수 없는 값이다. 그림 2 처럼 군집이 완전히 떨어진게 아닌 겹치는 부분이 있기 때문에 클러스터링이 결과 score가 그리 높지 않은 것으로 분석된다.

3.3.2 the score change with different K={2, 3, ..., 10}

클러스터의 개수가 변할 때, rand index 값이랑 silhouette score 값의 변화를 추가적으로 살펴보았다. 그 결과는 Table 3과 같다. rand index와 silhouette score은 K=3일 때, 엇비슷한 값을 갖지만 K 값이 증가함에 따라 값이 감소하게 된다. 이는 군집이 데이터의 원래의 군집 개수인 두 개 이상으로 형성이 되기 힘든 것을 뜻한다.

K	rand index	silhouette score
2	0.4914	0.6972
3	0.5174	0.6660
4	0.4127	0.5335
5	0.3418	0.5102
6	0.3210	0.4889
7	0.2620	0.4686
8	0.2599	0.4682
9	0.2345	0.4718
10	0.2310	0.4678

Table 3. K-means Scores Table

3.3.3 Discussion - comparing kmeans result with the classification results

단순히 feature만 가지고 군집을 분류하는 kmeans 클러스터링은 classification에 비해 성능이 좋지않다. 특히, WDBC 유방암 진단 데이터 셋은 그림 2와 같이 또렷한 공백 없이 두 군집이 겹쳐지는 부분이 존재하기 때문에 안좋았던 것으로 판단된다. 아마 두 군집이 겹쳐지는 부분이 없고 군집 사이의 거리가 멀다면 클러스터링 결과가 classification 결과와 비슷하지 않을까 예상한다.

4. Conclusion

2개의 classification 기법 (Random Forest, SVM)와 1개의 클러스터 기법(kmeans)에 대한 프로그램을 제작해보고, WDBC의 유방암 진단 데이터 셋을 이용하여 제작한 프로그램을 테스트 해보았다. classification의 경우, 해당 데이터가 sklearn 라이브러리에서 제공하는 정제된 예제 데이터기 때문에, miss value를 채우는 것과 같은 별도의 전처리 과정이 없었으며 결과 값 accuracy도 매우 높은 편으로 기록되었다. 반면에 클러스터링의 경우, 해당 데이터가 완전히 분리된 군집이 아닌 겹치는 부분이 있기 때문에 결과가 준수하지 못했다고 생각한다. 이번 과제를 하면서 머신러닝 프로그램 제작과 관련하여 좋은 경험이 되었다. 이번의 경험을 바탕으로 추후의 연구 과제에 머신러닝 기법을 적용해볼 예정이다.