**Machine Learning을 이용한**

**우울증 환자 분류 연구**

인천대학교 공과대학

전자공학과

장수민 (201701330)

2022년 11월 11일

목 차

**요약** --------------------------------------------------------------------------------

**Ⅰ. 서론** --------------------------------------------------------------------------

1.1. 연구 동기 ------------------------------------------------------------------

1.2. 연구 목표 ------------------------------------------------------------------

**Ⅱ. 본론** ----------------------------------------------------------------------------------------

2.1. HRV & ECG ----------------------------------------------------------------

2.2. 머신 러닝 모델 구현 -----------------------------------------------------

2.2.1. 교차 검증 (Leave-one-out) -------------------------------------------

2.2.2. 데이터 전처리 ----------------------------------------------------------

2.2.3. 모델 설정 (SVM) -------------------------------------------------------

2.2.4. 특성 선택 (SVM-RFE) -------------------------------------------------

2.3. 데이터 분석 ----------------------------------------------------------------

2.3.1. 하이퍼 파라미터 C 값 수동 조작 -------------------------------------

2.3.2. 유효한 특성 선택 수와 정확도 ----------------------------------------

2.3.3. 선택된 특성 ------------------------------------------------------------

**Ⅲ. 결론** ----------------------------------------------------------------------------------------

**참고문헌** --------------------------------------------------------------------------

**요 약**

우울증과 밀접한 생체 지표인 심박 변이도 관련 생체 지표들을 머신 러닝, 즉 기계 학습을 통해 우울증의 여부를 분류하고, 분류 정확도를 향상시킨다.

**Ⅰ. 서론**

**1.1. 연구 동기**

우울증 환자 분류에는 많은 생체 지표들이 관련되어 있다. 하나의 생체 지표가 우울증의 여부를 지배적으로 좌우할 수도 있고, 여러 생체 지표들의 조합이 좌우할 수도 있으며, 개인차가 있을 수도 있다. 그만큼, 우울증은 개개인의 특성을 모두 고려하여 정확히 진단하기 어려운 정신적 질환이다. 측정된 생체 지표feature를 머신 러닝machine learning 모델에서 활용하여 우울증 환자인지 아닌지를 분류하고, 분류 정확도를 향상시키기 위해 위 연구를 진행하게 되었다.

**1.2. 연구 목표**

나이와 성별,심박 변이도HRV, 그리고 우울증세의 여부를 담고 있는 csv 파일 형식의 정형데이터를 가공하고, 최적의 머신 러닝 모델을 구현하여 가장 높은 분류 정확도를 얻는 것이 목표이다.

**Ⅱ. 본론**

**2.1. HRV & ECG**

기본적으로 우울증 진단을 할 때, 심장박동 변이도HRV를 중요한 지표로 사용한다. HRV를 파악하기 위해 심전도ECG를 측정해야 한다. ECG는 심장이 박동할 때마다 발생하는 미약한 전기가 체내로 흐르고, 그 때의 신체 표면의 전위 분포를 도식화한 것이다. ECG 신호에서 1분 동안의 심장박동 횟수를 나타내는 심장 박동수를 유추할 수 있게 되며, 심장 박동수 1주기에 대한 미세한 변화의 정도를 HRV라 할 수 있다. HRV가 높으면 신경계가 스트레스와 회복 사이에 균형을 이루고 있다는 것을 보여주는 청신호이고, 낮으면 스트레스에 대한 내성이 낮아 운동 후의 회복이 제대로 이뤄지지 않고 있음을 암시한다. ECG와 HRV는 상기의 이유로 우울증 환자 분류에 사용할 수 있다.

**2.2. 머신 러닝 모델 구현**

**2.2.1. 교차 검증 (Leave-one-out)**

데이터 편중을 막기 위해서 별도의 세트로 구성된 학습 데이터 세트와 검증 데이터 세트에서 학습과 평가를 수행하는 검증 방식을 교차 검증Cross Validation이라 일컫는다. 데이터셋의 크기가 절대적으로 작은 점을 고려하여, 교차 검증 방식 중 하나인 Leave-one-out을 적용하였다. Leave-one-out 교차 검증은 크기가 n인 데이터셋을 n-1 크기의 훈련 데이터와 1의 크기를 갖는 테스트 데이터로 나누어 모델을 만든다.

**2.2.1. 데이터 전처리**

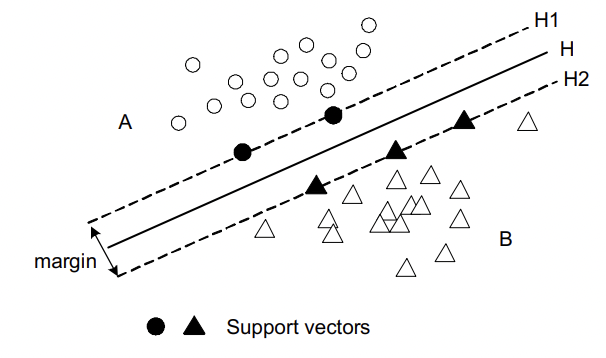
서로 다른 변수의 값 범위를 일정한 수준으로 맞추는 특성 스케일링feature scaling 작업을 위해 StandardScaler 혹은 MinMaxScaler 클래스를 사용하여 학습을 진행하였다. 하지만 특성 스케일링을 적용하지 않을 때, 더 좋은 성능을 나타냈다. 또한 총 102개의 특성 중 나이와 성별 특성을 제외했을 때 머신 러닝 모델의 성능이 향상되어, 해당 특성들도 제외하였다.

*StandardScaler() : 개별 feature를 평균이 0이고 분산이 1인 값으로 표준화*

*MinMaxScaler() : 데이터 값을 0과 1사이의 범위로 조정*

**2.2.2. 모델 설정 (SVM)**

MLPClassifier, KNN, SVM, RandomForestClassifier, LinearRegression, DecisionTreeClassifier, GaussianNB 분류 모델을 순차적으로 적용하여 정확도를 비교하고, 가장 좋은 모델을 선별하였다. SVMSupport Vector Machine은 대표적인 분류 알고리즘 중 하나로, 데이터를 두 종류의 결과 범주에 나누어 포함시켜야 하는 이진 데이터 분류에 좋은 성능을 보이므로, 위 연구에 적합하다. SVM은 데이터를 분류하는 초평면hyperplane을 구하는 알고리즘으로서, 초평면을 생성하는데 기여하는 서포트 벡터만 골라서, 초평면과 서포트 벡터 간의 거리가 최대가 되는 초평면을 긋는 방법을 찾아낸다. 선형 초평면에 의한 이항 분류만을 다루는 Linear SVM을 사용하여 Overfitting이나 Underfitting의 가능성을 최소화하였다.



**2.2.3. 특성 선택 (SVM-RFE)**

데이터를 가장 잘 판별하기 위해 선택적으로 특성을 선택하는 방법 중 회귀적 특성 제거RFE, Recursive Feature elimination를 사용하여 학습을 진행하였다. 비용 함수에서의 변화량을 순위의 기준으로 사용하고, 가장 작은 기준 값을 갖는 특성을 재귀적으로 제거하는 것을 의미한다.

*비용 함수의 변화량 :*

(단계 1) 비용함수에 대하여 가중치 를 최적화

(단계 2) 모든 특성에 대해 순위를 계산

(단계 3) 가장 낮은 순위 값을 갖는 특성 제거

(단계 4) (단계 1) – (단계 3) 반복

Guyon 등의 연구에서는, RFE 알고리즘 내의 가중치를 SVM의 해의 절댓값으로 사용하는 SVM-RFE를 제시한다. 본 연구에서는 SVM-RFE 알고리즘을 사용하여 우울증 환자 분류를 진행한다.

**2.3. 데이터 분석**

**2.3.1. 하이퍼 파라미터 C 값 수동 조작**

선형SVM 모델에서 사용자가 지정할 수 있는 입력 값, 즉 하이퍼 파라미터로는 C가 있다. 데이터의 차원이 높은 경우 선형적으로 완벽하게 분리해내는 것이 거의 불가능한데, 이를 해결하기 위해서 오류를 허용하는 전략이 고안되었다. 이와 관련된 하이퍼 파라미터가 C이다. C의 값을 낮게 설정하면 이상치outlier들이 많이 있을 가능성을 감안하여 완벽한 선형에 가까운 일반적인 결정 경계, 즉 초평면을 구한다. C의 값을 높게 설정하면 이상치들의 존재 가능성을 희미하게 보고, 좀 더 세심한 결정 경계를 찾아내므로 완벽한 선형 경계에서 멀어지는 형태를 띤다.

데이터 특성에 따라 가장 좋은 성능을 나타내는 머신 러닝 모델도 다르지만, 하이퍼 파라미터 값도 다르다. 그래서 데이터마다 하이퍼 파라미터의 값을 찾는 과정이 필요한데 이를 하이퍼 파라미터 튜닝이라 한다. 보통 사이킷런에서 제공하는 모듈인 GridSearchCV나 RandomSearchCV를 사용하여 실험하고자 하는 하이퍼 파라미터와 값 목록을 지정하고, 가능한 모든 조합에 대해 학습을 수행한다. 하지만 본 연구에서는 학습 비용을 최소화하고, 튜닝해야 할 대상이 C에만 국한되는 점을 고려하여 직접 C의 값을 찾았다.

텍스트이(가) 표시된 사진

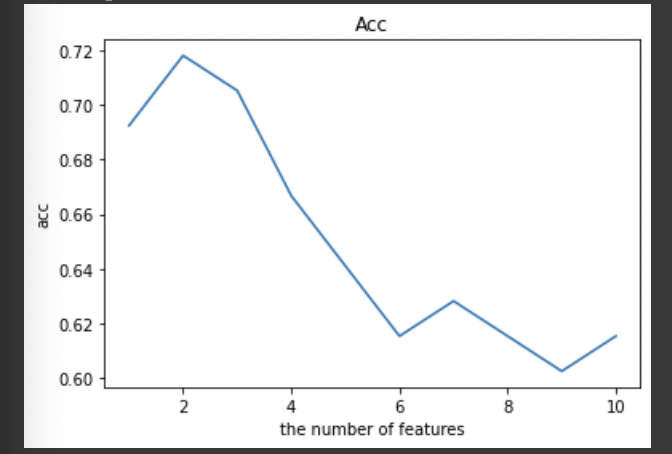
자동 생성된 설명

*0.01 ~ 1 의 값을 로그 단위로 분할*

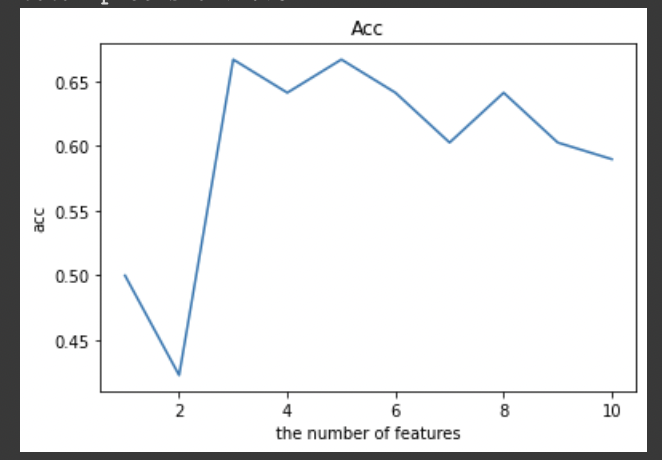
20개로 분할된 로그 단위 값을 사용하여 학습을 진행했을 때, 최적의 C의 값은 0.01 ~ 0.026에 근사하였고, 근사한 값 범위 내에서는 유의미한 성능 차이를 보이지 않아 0.01, 0.1, 1의 값을 학습에 사용하여 성능 추이를 확인하였다.

**2.3.2. 유효한 특성 선택 수와 정확도**

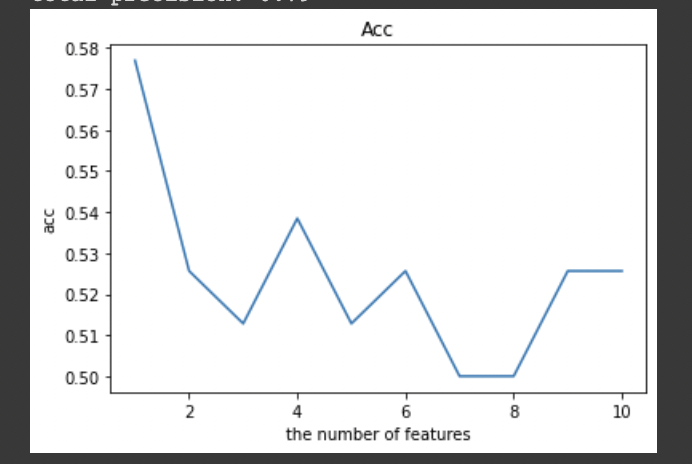
SVM-RFE를 적용하여 재귀적으로 특성을 제거하고 최종적으로 남길 특성 개수의 범위를 지정하기 위해 학습 성능의 결과로 대체로 정확도accuracy가 45% 내외, 혹은 그 이상인 결과에 대해 집중하였다. 보통 1개에서 10개의 특성을 선택하였을 때 그러한 결과를 얻을 수 있었다.



C = 0.01



C = 0.1



C = 1

**2.3.3. 선택된 특성**

csv형식의파일에는심박 변이도 관련 파라미터 20개를 5개의 단계(P1 : Baseline, P2 : Stress task, P3 : Recovery, P4 : Relaxation task, P5 : Recovery)에 걸쳐 측정한 총 100개의 특성 값이 있다. 심박 변이도 관련 파라미터 20개는 다음 목록의 내용과 같다.

텍스트이(가) 표시된 사진

자동 생성된 설명

학습의 결과로 얻은 선택된 특성은 다음의 표와 같다.

테이블이(가) 표시된 사진

자동 생성된 설명

**Ⅲ. 결론**

데이터 분석 과정 중에 과대적합으로 인해 비약적인 정확도의 향상이 일어나는 현상에 주목하여 모델을 변경하거나 하이퍼 파라미터를 조정하고, 최대한 불필요한 논리적 반복을 피해 학습 시간을 줄이고 동시에 학습 비용을 최소화하고자 하였다. 우울증 분별을 위해 구현한 본 연구의 머신 러닝 모델은 다층신경망 분류 모델Multi Layer Perceptron Classifier을 사용할 때보다 정확도가 향상되었고, 반복되는 논리의 횟수가 줄어들어 중첩 교차 검증Nested Cross Validation을 사용할 때보다 같은 환경 조건 내에서 학습 시간이 줄어들었다. 결과적으로C가 0.01이고 2개의 특성으로 학습을 진행했을 때 71.8%의 정확도Accuracy를 얻었고, 1개에서 10개까지의 특성을 선택하여 학습을 진행했을 때 대체로 80% 내외의 정밀도Precision를 얻을 수 있었다. 데이터셋의 절대적인 크기가 분류 정확도에 매우 지배적인 요인이나, 비교적 작은 규모이면서 다차원인 데이터셋에 적용할 수 있는 머신 러닝 모델을 구현했다는 점에서 의의가 있고, 데이터 특성과 규모에 따라 머신 러닝 모델은 더 심화되어야 할 것이다.

**참고문헌**

[1] scikit learn – 1.4. Support Vector Machines

[2] Fault Classification for Rotating Machinery Using Support Vector Machines with Optimal Features Corresponding to Each Fault Type – Yang Seok Kim, Do Hwan Lee and Seong Kook Kim

[3] Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." *Machine learning* 46.1-3 (2002): 389-422.

[4] Detection of major depressive disorder from linear and nonlinear heart rate variability features during mental task protocol – Sangwon Byun, Ah Young Kim, Eun Hye Jang, Seunghwan Kim, Kwan Woo Choi, Han Young Yu, Hong Jin Jeon