게놈 관련성 분류 흐름

이 문서는 게놈 관련성 분류 파이프라인의 데이터 준비부터 최종 보고까지의 단계별 프로세스를 설명합니다. 이 파이프라인은 cM_1, cM_3, cM_6의 세 가지 다른 센티모건 임계값을 처리하며 유사한 흐름을 가지지만 성능이 다릅니다.

1단계: 원시 데이터 수집

- **사용된 데이터 파일**:
  - `model_input_with_kinship_filtered_{dataset}.csv` (기본 레이블링된 게놈 데이터)
  - `merged_info.out` (기본 통계적 특징, 103개 열)
  - `merged_added_info.out` (추가 통계적 특징, 123개 열에 거리 임계값 포함)
- **프로세스**: 각 데이터셋 변형에 대해 세 개의 별도 데이터 소스 로드
- **결과**: 메모리에 로드된 원시 데이터 파일
- **통계**: CSV는 게놈 쌍 관계와 친족 레이블(1, 2, 3, 4, 5, 6, UN)을 포함하며, .out 파일은 IBD 분석에서 통계적 메트릭을 포함합니다.

2단계: 데이터 병합 및 특징 통합

- **사용된 데이터 파일**: 1단계의 세 개 파일 모두
- **프로세스**:
  - 두 .out 파일에서 통계적 특징 파싱 및 키:값 추출
  - 'pair' 식별자를 기준으로 기본 및 추가 통계 데이터 병합
  - 기본 값을 우선하고 추가 값으로 누락된 값 백필하여 열 충돌 처리
  - 게놈 레이블링 쌍 데이터와 통계적 특징 병합
- **결과**:
  - 친족 레이블과 통계적 특징이 포함된 단일 통합 데이터셋
  - 모든 필수 열 포함: pair, IBD1_len, IBD2_len, R1, R2, Num_Segs, Total_len, kinship
  - <1cM에서 <10cM까지의 거리 임계값에 대한 원시 개수 및 백분율을 포함한 통계적 특징
  - 출력 파일: `data/processed/merged_<dataset>.csv` (UN 포함 시나리오) 및 `data/processed/merged_<dataset>_noUN.csv` (UN 제거 시나리오)
- **통계**:
  - 기본 통계 열: 103개 ('pair' 제외)
  - 추가 통계 열: 123개 ('pair' 제외)
  - 기본과 추가 간 공통 열: 104개
  - 추가 전용 열: 20개
  - 최종 병합 데이터셋: 친족 레이블 외 124개 총 열

3단계: 레이블 처리 기반 시나리오 생성

- **사용된 데이터 파일**: 2단계의 병합 데이터셋
- **프로세스**: 'UN'(알 수 없음) 친족 레이블 처리에 따라 두 시나리오로 분할
- **결과**:
  - **포함 시나리오**: 'UN' 레이블을 포함한 모든 친족 레이블 유지 (cM_1/cM_3의 경우 2,805개 샘플, cM_6의 경우 2,787개 샘플)
  - **NoUN 시나리오**: 'UN' 레이블 제거 및 특정 친족 관계만 유지 (모든 변형에서 882개 샘플, 클래스: 1,2,3,4,5,6만)
- **통계**:
  - 포함 시나리오의 클래스 분포: 'UN'이 모든 샘플의 68-69% 차지
  - NoUN 시나리오 생성 시 제거된 샘플: 약 1,923개

    - 보존된 친족 클래스: 1, 2, 3, 4, 5, 6 (알 수 없는 관계 제외)

4단계: 특징 엔지니어링 및 선택

- **사용된 데이터 파일**: 3단계의 시나리오 데이터셋
- **프로세스**:
    - RandomForest 중요도 순위로 통계적 특징 선택 적용
    - 사용 가능한 통계 변수에서 가장 예측력이 높은 75개 특징 식별
    - 수학적 조합을 통한 추가 집계 특징 생성
    - 모든 수치 특징에 걸친 특징 스케일링 적용
- **결과**:
    - 각 시나리오에 대한 최적화된 특징 세트
    - 일관된 데이터 변환을 위한 훈련된 스케일러
    - 특징 중요도 순위 및 시각화
    - 출력 파일: 스케일러 파일 `data/processed/scaler_<dataset><suffix>.pkl` 로 저장, 상위 특징 `data/processed/top_features_<dataset><suffix>.pkl` 로 저장, 특징 중요도 플롯 `reports/<dataset>/assets/<scenario>/feature_importance_<dataset>_<scenario>.png` 및 `.svg`
- **통계**:
    - 선택된 특징: 사용 가능한 변수에서 상위 75개
    - 특징 유형: IBD 세그먼트, 거리 임계값, 비율, 백분율
    - 스케일러 유형: 모든 수치 특징에 걸친 표준화

5단계: 클래스 불균형 처리와 모델 훈련

- **사용된 데이터 파일**: 4단계의 엔지니어링 및 선택된 특징 데이터셋
- **프로세스**:
    - 세 가지 다른 모델 아키텍처 훈련:
        - **RandomForest**: 특징 중요도가 있는 전통적 앙상블 방법
        - **MLP**: 5개 레이어의 다층 퍼셉트론 신경망
        - **CNN**: 패턴 인식을 위한 1D 컨볼루션 신경망
    - 네 가지 클래스 불균형 처리 전략 적용:
        - **Zero**: 재균형 없음 (기준, 다수 클래스를 선호할 수 있음)
        - **Weighted**: 훈련 중 클래스 가중치 손실 함수
        - **SMOTE**: 합성 소수 오버샘플링 기법
        - **OverUnder**: SMOTE 오버샘플링 + ENN/Tomek 언더샘플링 결합
- **결과**:
    - 데이터셋 변형당 24개 훈련된 모델 (2개 시나리오 × 4개 모드 × 3개 모델)
    - 모델 가중치 및 구성 파일
    - 훈련 진행 및 성능 로그
    - 출력 파일: 훈련된 신경망 모델 `models/<dataset>/<scenario>/<imbalance_mode>/<model>.pth` 로 저장 (MLP/CNN), 훈련 메타데이터 `models/<dataset>/<scenario>/<imbalance_mode>/training_meta.json`
- **통계**:
    - 훈련 기간이 크게 다름:
        - Zero 모드: 138.8-243.7초
        - Weighted 모드: 138.9-252.4초
        - OverUnder 모드: 197.1-244.7초

        - SMOTE 모드: 963.4-1,071.6초 (합성 데이터 생성으로 인해 가장 김)
    - 모든 훈련이 CUDA 디바이스에서 효율성을 위해 수행
    - 모델당 100개 훈련 에포크

6단계: 포괄적 모델 평가

- **사용된 데이터 파일**: 5단계의 모든 훈련된 모델, 검증 데이터셋
- **프로세스**:
    - 여러 성능 메트릭을 사용하여 각 모델 평가
    - 분류 분석을 위한 혼동 행렬 생성
    - 추가 분석을 위한 확률 예측 내보내기
    - 시나리오, 모드 및 모델 유형에 따른 성능 비교
- **결과**:
    - 변형당 24개 모델 조합에 대한 성능 순위
    - 시각화가 포함된 상세한 혼동 행렬
    - 신뢰도 점수 및 확률 예측
    - 비교 분석 테이블 및 차트
    - 출력 파일: 성능 결과 `reports/<dataset>/results.json`, 상세 보고서 `reports/<dataset>/results.md`, 혼동 행렬 플롯 `reports/<dataset>/plots/confusion/<scenario>/<imbalance_mode>/`, 확률 예측 `reports/<dataset>/assets/<scenario>/probabilities_<model>_<imbalance_mode>.json`
- **통계**:
    - **포함 시나리오 최고 성능**:
        - cM_6: F1-weighted 0.9555, AUC-weighted 0.9962 (RandomForest overunder)
        - cM_1: F1-weighted 0.9460, AUC-weighted 0.9941 (RandomForest overunder)
        - cM_3: F1-weighted 0.9420, AUC-weighted 0.9928 (RandomForest overunder)
    - **NoUN 시나리오 최고 성능**:
        - cM_1: F1-weighted 0.9234, AUC-weighted 0.9905 (RandomForest overunder)
        - cM_6: F1-weighted 0.9141, AUC-weighted 0.9874 (RandomForest overunder)
        - cM_3: F1-weighted 0.8929, AUC-weighted 0.9839 (RandomForest overunder)
    - **일관된 패턴**: 모든 변형에서 RandomForest가 MLP와 CNN보다 우수
    - **검증 샘플 크기**: 포함의 경우 ~697-702개, NoUN의 경우 ~221개

7단계: 자동화된 보고서 생성

- **사용된 데이터 파일**: 6단계의 평가 결과, 모델 통계, 특징 중요도 데이터
- **프로세스**:
    - 실행 요약이 포함된 포괄적 마크다운 보고서 생성
    - 혼동 행렬 플롯 및 성능 차트와 같은 시각화 생성
    - 접근성을 위한 영어 및 한국어 이중 보고서 생성
    - 공유 및 보관을 위한 PDF 버전 생성
    - 모든 변형에 걸친 성능 비교 테이블 컴파일
- **결과**:
    - 상세한 분석 및 통찰력이 포함된 전문 보고서
    - 모델 성능 비교를 보여주는 시각적 대시보드
    - 특징 중요도 플롯 및 분석
    - 방법별 훈련 시간 및 계산 효율성 메트릭
    - 출력 파일: 이중 언어 보고서 `reports/<dataset>/results.md` 및 `reports/<dataset>/results_KR.md`, 시각화 자산 `reports/<dataset>/assets/`, 플롯

`reports/<dataset>/plots/`, 테이블 `reports/<dataset>/tables/`, PDF 버전

- **통계**:
  - 보고서는 정확도, F1 (weighted/macro), AUC (weighted/macro/micro)와 같은 상세 메트릭 포함
  - 다른 불균형 처리 방법에 걸친 훈련 시간 비교
  - GPU 활용 및 계산 효율성 분석
  - cM 변형 (cM_1, cM_3, cM_6)에 걸친 비교 성능 분석

cM 변형 간 변동성

성능 특성

- **cM_6**: 포함 시나리오에서 최고 전반적 성능, 가장 계산 효율성 높음
- **cM_1**: NoUN 시나리오에서 최고 성능, 균형 잡힌 효율성
- **cM_3**: 중간 성능, SMOTE 모드에서 가장 긴 훈련 시간

계산 패턴

- **훈련 효율성 순위**: 대부분 불균형 모드에서 cM_6 > cM_1 > cM_3
- **SMOTE 모드**: 모든 변형에서 4-5배 긴 훈련 시간 일관되게 요구
- **GPU 활용**: 모든 변형이 신경망 훈련 중 효율적인 CUDA 활용 표시

데이터 구조 일관성

- 모든 변형이 동일한 전처리 단계 및 데이터 구조 따름
- 샘플 크기 변동이 최소 (cM_6이 약간 적은 샘플 보유)
- 특징 엔지니어링 프로세스가 모든 변형에 걸쳐 표준화
- 모델 아키텍처 및 하이퍼파라미터가 일관되게 유지

요약

게놈 관련성 분류 파이프라인은 특징 엔지니어링, 모델 훈련 및 평가의 여러 단계에서 복잡한 게놈 데이터를 성공적으로 처리하는 정교한 머신러닝 시스템을 나타냅니다. 이 시스템은 다른 센티모건 임계값에 걸쳐 강력한 성능을 보여주며, cM_6은 알 수 없는 관계를 포함한 친족 분류 작업에 가장 유망하고, cM_1은 알 수 없는 관계가 훈련에서 제외될 때 최고 결과를 제공합니다.