

06

RPI를 통한 신약후보물질 예측 및 추천 머신러닝 기법 개발

소속 정보컴퓨터공학부

분과 A

팀명 졸업가능하조

참여학생 이지현, 박성아, 손우정

지도교수 송길태

과제 개요

과제 선정 배경

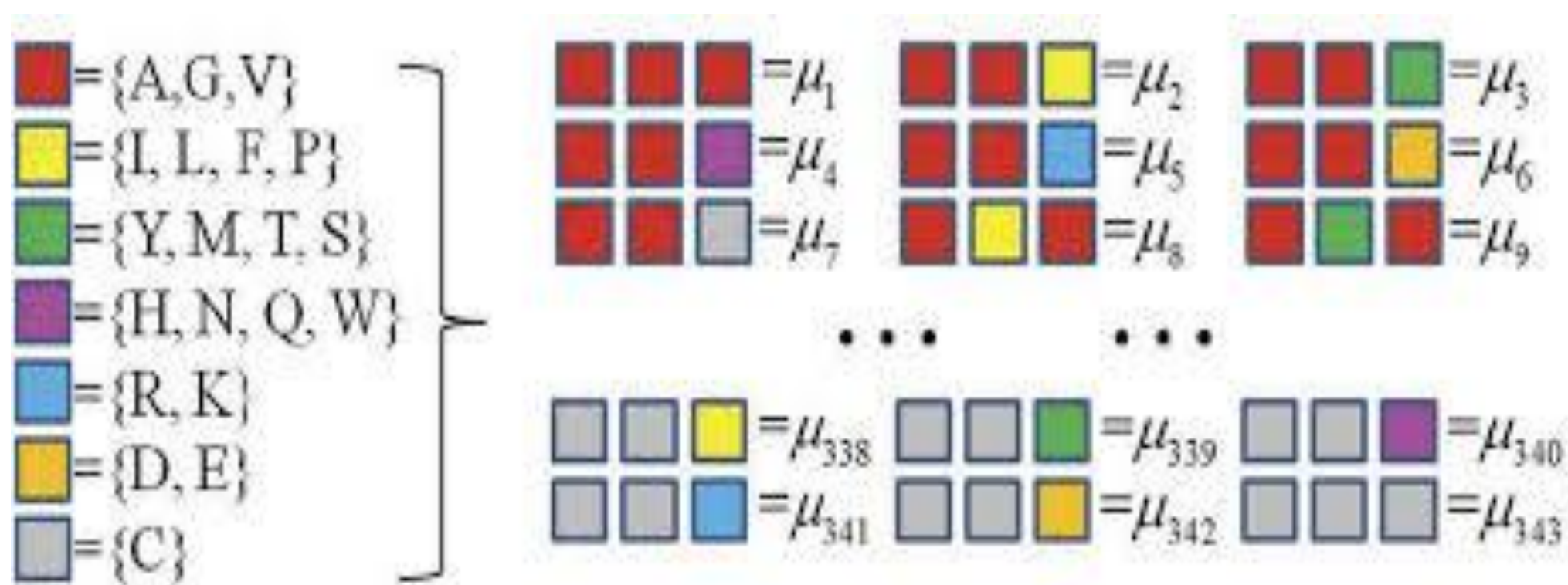
Bioinformatics

- 생물학, 분자 생물학 등의 실험에서 컴퓨터 과학을 응용하여 분자 수준에서 분석, 연구, 학문
- 수작업으로 처리하기 힘든, 많은 실험 데이터 분석을 위해 컴퓨터 과학의 알고리즘 등을 활용하는 연구, 학문 분야

과제 목표

- 20개의 아미노산으로 구성되는 Protein 서열과 4개의 핵산으로 구성되는 RNA 서열이 주어졌을 때 결합 유무를 분류하는 머신러닝 기법을 개발한다
- Classifier의 성능을 높이는 프로젝트를 진행한다

시스템 개요



사용된 Classifier

- Random Forest : 여러 개의 Decision Tree를 조합해 사용하는 Ensemble 알고리즘
- Support Vector Machine : 분류를 위한 기준 선을 정의한다. 경계의 어느 쪽에 속하는지 분류
- Gradient Boost : 가중치 업데이트를 경사하강법을 이용해 최적화된 결과를 얻음
- XGBoost : 기존 GBM의 속도 문제를 해결하기 위해 전산 속도와 모델 성능에 초점을 맞춤
- AdaBoost : 약한 학습기의 오류데이터에 가중치를 부여하며 부스팅을 수행
- Bagging : 동일한 알고리즘으로 여러 분류기를 만들어 보팅으로 최적 결정
- Soft Voting : 각 Class별로 모델들이 예측한 probability를 합산해 가장 높은 class 선택

기대효과

- 실험 시간이 너무 길어 비용이 과도하게 투입되는 생물학 실험을 구조 결합 데이터를 바탕으로 실험 비용을 줄일 수 있다.
- 정확도가 특히나 중요한 생물학 분야에서 기존보다 성능을 높인 Classifier로 보다 정확한 실험 결과 예측이 가능하다.