Využitie strojového učenia na predikciu vplyvu mutácií na stabilitu proteínov

Juraj Ondrej Dúbrava

Brno University of Technology, Faculty of Information Technology
Božetěchova 1/2. 612 66 Brno - Královo Pole

xdubra03@fit.vutbr.cz

January 26, 2018

Motivácia

- Vplyv mutácií na stabilitu
- Neuspokojivé výsledky predikčných nástrojov
- Zlepšenie predikcie - návrh stabilnejších proteínov, účinnejších liečiv,...

Proteíny

- základné stavebné prvky všetkých organizmov
- zabezpečujú množstvo funkcií
- proteín je tvorený reťazcom aminokyselín
- vlastnosti proteínov sú ovplyvnené mutáciami

Stabilita proteínov

- jedna z dôležitých vlastností proteínov, súvisí so stavom proteínu
- dôležitosť skúmania stability pre rôzne oblasti
- stabilné proteíny - lepšie zvládnutie nepriaznivých okolitých podmienok, vysokých teplôt,...
- pôsobenie mutácií na stabilitu, stabilizujúce vs. destabilizujúce mutácie
- snaha o predikciu ich vplyvu, využitie strojového učenia

Dataset

- základ pre strojové učenie
- 1564 záznamov mutácií - 1255 destabilizújúcich, 309 stabilizujúcich
- 7 parametrov datasetu
- výber metódy strojového učenia
- prvotné testovanie - nástroj WEKA, odskúšanie ponúkaných algoritmov
- najlepšie výsledky - metódy Random Forest, SVM

Dataset

Metóda	TP rate	FP rate	Accuracy
Naive Bayes	0,784	0,719	0,765
LibSVM	0,786	0,706	0,766
SMO	0,774	0,774	0,6
DecisionTable	0,774	0,774	0,6
RandomForest	0,793	0,692	0,797
RandomTree	0,793	0,574	0,766
J48	0,774	0,626	0,74

Implementácia

- spôsob implementácie - Python skript
- použitie knižnice Scikit-learn na implementáciu algoritmov strojového učenia
- prvá fáza - odskúšanie metódy Random Forest
- neuspokojivé výsledky samostanej metódy
- implementácia a odskúšanie metódy SVM

- vylepšenie predikcie - použitie ensemble systému
- spojenie klasifikátorov - metódy Random Forest a SVM
- použitie pri nedostatku dát, realistickejší výsledok
- vytvorenie menších trénovacích podsád, bagging
- vylepšenie výpočtu parametra konzervovanosti - Felensteinov algoritmus
- presnosť predikcie - približne 74%, korelácia 0.32

Záver

- analýza dátovej sady a jej použitie pre strojové učenie
- implementácia skriptu na výpočet parametrov
- implemtácia a otestovanie metód Random Forest a SVM
- použitie ensemble stratégie na vylepšenie presnosti
- výhoda rýchlej a dostatočne presnej predikcie
- možnosť použitia pre vytipovanie zaujímavých mutácií na ďalšie skúmanie

Ďakujem za pozornosť !