Zarovnávanie sekvencií s použitím metód klasifikácie

Diplomová práca

Bc. Michal Hozza

Vedúci práce: Mgr. Tomáš Vinář, PhD.
Konzultant: Mgr. Michal Nanási

Fakulta matematiky, fyziky a informatiky, Univerzita Komenského, Bratislava

18. marca 2013

Obsah

Cieľ

- Cieľom práce je vytvoriť nové metódy na korekciu zarovnaní biologických sekvencií na základe prídavnej informácie.
- Integrácia tejto informácie bude zabezpečená pomocou techník využívaných na klasifikáciu v strojovom učení.

Zarovnávanie sekvencií

Kľúčové problémy:

- Aké typy zarovnávania by sme mali uvažovať
- Skórovací systém, ktorý použijeme na ohodnotenie zarovnania a trénovanie
- Algoritmus, ktorý použijeme na hľadanie optimálneho alebo dobrého zarovnania podľa skórovacieho systému
- Štatistická významnosť zarovnania.

Modely

Generatívny:

- sa snaží modelovať proces, ktorý generuje dáta ako pravdepodobnosť $P(X, Y, Z)$
- rozložíme ju pomocou nezávislých predpokladov na procese $\longrightarrow$ obmedzujúce

Modely

Generatívny:

- sa snaží modelovať proces, ktorý generuje dáta ako pravdepodobnosť $P(X, Y, Z)$
- rozložíme ju pomocou nezávislých predpokladov na procese $\longrightarrow$ obmedzujúce

Diskriminačný

- priamo odhaduje $P(Z|X, Y)$ alebo prislúchajúcu diskriminačnú funkciu, a preto sa zamerá na podstatnú časť problému odhadu
- Nepotrebuje nezávislosť $\longrightarrow$ silnejšie

Existujúce riešenia

- Problém inverzného zarovnania

Existujúce riešenia

- Problém inverzného zarovnania
- Support vector training of protein alignment models
  - Support Vector Machine (SVM)
  - Umožňuje trénovať pomocou rôznych účelových funkcií

Existujúce riešenia

- Problém inverzného zarovnania
- Support vector training of protein alignment models
  - Support Vector Machine (SVM)
  - Umožňuje trénovať pomocou rôznych účelových funkcií
- Contralign: Discriminative training for protein sequence alignment.
  - Conditional Rnadom Fields (CRF)
  - Neumožnuje trénovať pomocou rôznych účových funkcií

Odlišnosti nášho riešenia

- Korekcia existujúcich zarovnaní
  - Použitie súbežne s existujúcimi zarovnávačmi
- Rôzne metódy trénovania
- Možnosť učenia bez učiteľa
- Iný klasifikátor
  - možno porovnanie viac rôznych klasifikátorov
  - prípadne abstrakcia od klasifikátora

Klasifikátor

Random Forest

- Zložený z klasifikačných (rozhodovacích) stromov
- Stromy hlasujú o výsledku ktoré hlasujú

Simulátor

- Model určený na prvotné experimenty
- Program simuluje evolúciu
  - Generovanie dvojice postupností so správnym zarovnaním
  - Generovanie dodatočej informácie
  - Simulácia mutácie a delécie

Ďakujem za pozornosť!