Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Matyáš Lamprecht
Název práce	Modelování molekulární podobnosti pomocí fragmentů
Rok odevzdání	2019
Studijní program	Informatika
Studijní obor	Obecná informatika

Autor posudku	RNDr. František Mráz, CSc.	Oponent
Pracoviště	Katedra softwaru a výuky informatiky	

K celé práci

	lepší	OK	horší	nevyhovuje
Obtížnost zadání		X		
Splnění zadání		X		
Rozsah práce *... textová i implementační část, zohlednění náročnosti*		X		

Práca sa zaoberá navrhovaním metód pre tzv. virtuálny screeneing, kde na základe chemického popisu molekuly a známej aktivity trénovacej sady molekúl, chceme predikovať aktivitu nových neznámych molekúl. Autor navrhol a vyskúšal niekoľko metód na vylepšenie predikcie biologickej aktivity molekuly na základe podobnosti molekuly (presnejšie popisu vlastností celej molekuly a vlastností fragmentov molekuly) so známymi biologicky aktívnymi molekulami. Jedná sa vlastne o úlohu strojového učenia. Veľmi pozitívne hodnotím návrh metód na združovanie hashovaných deskriptorov nazývaných indexy do skupín s rôznymi metódami vyhodnocovania podobnosti sád takýchto deskriptorov.

Tiež je treba vyzdvihnúť dôkladné testovanie metód zavŕšené dôslednou analýzou dosiahnutých výsledkov. Dosiahnuté výsledky sú zaujímavé, ale ich kvalitu kazia nedostatky v popise metód výpočtov podobnosti, viz nižšie.

Textová část práce

	lepší	OK	horší	nevyhovuje
Formální úprava *... jazyková úroveň, typografická úroveň, citace*		X		
Struktura textu *... kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu*		X	X	
Analýza		X		
Vývojová dokumentace		X		
Uživatelská dokumentace		X	X	

Práca je napísaná s minimom gramatických chýb (napr. str. 17, r. 3: chýbajúca čiarka za slovom "argumenty" vo vete "..., kde ke každému skriptu jsou uvedeny všechny argumenty je v příloze A.1."; str. 35, r. −1: preklep "jedné aktivní molekulu"). Inak má adekvátnu štruktúru i obsah. Avšak popis metód je veľmi stručný. U niekoľkých vzorcov chýba popis premenných a zvolené značenie veľmi sťažuje pochopenie textu.

- Totožné značenie sa používa na niekoľko rôzne počítaných hodnôt. Napr. na strane 9 vzorce (1.1), (1.2) a (1.3) definujú tri rôzne podobnostné koeficienty, ale označujú ich rovnako S_{AB} a v texte po (1.3) sa dokonca tvrdí, že definícia S_{AB} je vo všetkých troch zhodná.

- Vo vzorci (1.4) pre "enrichment factor" na str. 11 chýba popis všetkých premenných (θ, r_i, n, N); text, ktorý nasleduje, popisuje, aké je problematické nastavovanie percentuálneho podielu molekúl, ktoré chceme dostať z virtuálneho screeningu. Prečo sa nestanoví rovno počet molekúl, ktoré chceme dostať?

- str. 13, r. 10: veta "Jednotlivé experimenty jsou implementovány modely." spomína modely. Ďalej v texte sa hovorí tiež o modeloch, ale nerozlišuje sa medzi matematickým modelom a triedou `IModel` z implementácie.

- str. 13, r. −13: z popisu "...se vezme bitový vektor molekuly a pozice bitů se vymodulí číslem n" nie je jasné, čo je výsledkom – je to pozícia nulových i nenulových bitov, nový kratší bitový vektor (ako sa nastavia bity, keď sa jednu pozíciu má namapovať viacej bitov s rôznymi hodnotami), alebo niečo úplne iné?

- str. 30, r. 16–17: "Pokud budeme dávat všechny aktivní molekuly dohromady, budeme je značit A_L." – A_L bude množina molekúl, alebo množina indexov, o ktorých sa hovorí vo vete, ktorá je pred touto, alebo to bude multimnožina indexov?

- str. 32, oddiel 5.1.4 uvádza príklad výpočtu skóre pre "deskriptor model", ale výpočet sa nedá overiť, pretože nie je jasné, ktoré deskriptory boli použité pri výpočte hodnôt vo vzorcoch (5.6) a (5.7).

- str. 34, vzorec (5.13): čo je to a_i? Ak je to jeden index, tak čo znamená $|a_i|$? Keď A_P označuje počet aktívnych molekúl, prečo je vo vzorci použitá absolútna hodnota z A_P?

- str. 34, r. −4: kryptická schéma "$[a, b, c] -> [[a, b], c]$" bez popisu nestačí na vysvetlenie, čo sa deje s množinou indexov.

- str. 35, oddiel 5.3.3: popis hovorí, že odstránime index 4089138501, ale index 3624155 by mal byť zachovaný spolu so skupinou [3624155,4089138501], teda veľkosť $|A_1|$ by mala byť 21. Text uvádza, že veľkosť $|A_1|$ je 20, teda by sa odstránil aj index 3624155, čo z textu nevyplýva.

- Z textu nie je jasné, keď čo sa deje, keď sa tvorba skupín robí iterovane. Pri druhej iterácii vznikajú skupiny maximálnej veľkosti 3 alebo maximálnej veľkosti 4?

V užívateľskej dokumentácii sú tiež miestami vynechané informácie. Napr. príklad v oddieli 3.3 na str. 19 využíva implicitný adresár pre dáta a popis hovorí, že keby sa mali použiť iné sady molekúl, tak by sa k nim museli nastaviť cesty. Ako sa majú cesty nastaviť, už nie je uvedené. V skutočnosti všetky skripty používajú jednotné nastavovanie ciest, ale užívateľ sa to dozvie až zo zdrojových kódov.

Implementační část práce

	lepší	OK	horší	nevyhovuje
Kvalita návrhu *... architektura, struktury a algoritmy, použité technologie*		X		
Kvalita zpracování *... jmenné konvence, formátování, komentáře, testování*		X		
Stabilita implementace	X			

Priložené programy sú plne funkčné a umožňujú jednoducho robiť počítačové experimenty i generovať grafy. Programy sú nedostatočne komentované. Komentáre sú spravidla iba na začiatku zdrojového textu modulu, ale chýbajú u jednotlivých funkcií a tried.

Architektúra aplikácie, tj. sada funkcií a modulov je navrhnutá veľmi prehľadne.

Pri svojich experimentoch som narazil iba na jediný drobný problém. Keď predikcia zlyhala a vrátila všetkým testovacím molekulám totožnú aktivitu, tak výsledkom bola skvelá hodnota AUC 1.0 a vysoké hodnoty EF1 a EF2. Je to spôsobené triediacim algoritmom a tým, že všetky aktívne molekuly boli v testovacej sade pohromade na konci.

Celkové hodnocení Velmi dobře

Práci navrhuji na zvláštní ocenění Ne

Datum 31.8.2019 Podpis