

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	Kompenzace vlivu matoucích faktorů při klasifikaci RNASeq dat
Jméno autora:	Karina Balagazova
Typ práce:	bakalářská
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra kybernetiky
Oponent práce:	Mgr. Petra Hrubá, PhD.
Pracoviště oponenta práce:	Transplantační laboratoř IKEM

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání	náročnější
<i>Hodnocení náročnosti zadání závěrečné práce.</i>	
Vložte komentář.	

Splnění zadání	splněno s menšími výhradami
<i>Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.</i>	
<p><i>V úvodu práce studentka stručně popsala základní dogma molekulární biologie, genové exprese a metody RNA sekvenování. Dále se zabývá možnostmi kompenzace matoucích faktorů. A to jednak jejich odstranění pomocí designu studie (ideálně randomizované či case-control studie), to však v případě našeho projektu použít nelze. Poté popsala možnosti kompenzace matoucích faktorů pomocí statistických metod (stratifikací, regrese či propensity score matching). V samotné práci byly použity tři druhy klasifikátorů (SVM, regrese a kNN), vygenerovaná umělá data a na nich se studentka pokusila zjistit optimální počet potřebných pacientů ve skupině pacientů se stabilní funkcí, optimální počet příznaků a opakování, aby nedocházelo k překompensaci dat. Vždy provedla klasifikaci dat, jejich kompenzaci a klasifikaci umělých dat. Zadání práce je velmi aktuální, vzhledem k stále narůstajícímu počtu transkriptomických studií, které pracují s daty obsahující mnoho příznaků ve skupinách s malým počtem vzorků.</i></p>	

Zvolený postup řešení	správný
<i>Posuďte, zda student zvolil správný postup nebo metody řešení.</i>	
Vložte komentář.	

Odborná úroveň	C - dobře
<i>Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.</i>	
Vložte komentář.	

Formální a jazyková úroveň, rozsah práce	C - dobře
<i>Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.</i>	
<p><i>Práci vypracovala zřejmě studentka, jejímž rodným jazykem není čeština, a proto obsahuje velkou řadu stylistických chyb. Bohužel, některé chyby při překladu vedly i k faktickým chybám a nepřesnostem. Např. nukleotid není základní stavební jednotkou buňky, ale DNA (kapitola 1.1). Na straně 1 "Operační tolerance je vzácný jev, vyskytující se u některých pacientů, kteří z různých důvodů přestali užívat imunosupresivní léky poté, co podstoupili transplantaci ledvin." Správně by mělo být: Operační tolerance je stav, kdy pacienti po transplantaci ledvin z různých důvodů přestanou brát imunosupresi po dobu nejméně jednoho roku, přesto nedojde k poškození funkce štěpu". Porovnávání (matching) by mělo být asi správně přeloženo jako párování. Obr.4.14. a další by neměl být označen pouze jako graf č.1, ale vysvětleno co je na něm</i></p>	

znázorněno a jakým způsobem byl získán (liší se parametry IS effectu a to by mělo být vysvětleno), pacientů ve skupině CR bylo pouze 14, ne 15.

Výběr zdrojů, korektnost citací

B - velmi dobře

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Vložte komentář.

Další komentáře a hodnocení

Vložte komentář (nepovinné hodnocení).

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

Předkládaná bakalářská práce se zabývá způsoby kompenzace matoucích faktorů při klasifikaci dat z RNA sekvenování. Data ze sekvenování poskytnutá z projektu IKEM, byla získána od pacientů po transplantaci ledviny, kteří mají tzv. operační toleranci. To je situace, kdy přestanou z různých důvodů brát imunosupresivní léky po dobu nejméně jednoho roku, které potlačují odhojování ledvinného štěpu, a přesto jim ledvina funguje. Tento jev je však velmi vzácný s incidencí asi 3/10 000 případů. Identifikace pacientů, u kterých by bylo možné imunosupresi minimalizovat, by byla velmi přínosná. Cílem projektu bylo tedy pokusit se identifikovat pacienty s operační tolerancí pomocí několika pečlivě vybraných kontrolních skupin. Z kontrolních skupin byla nejdůležitější skupina pacientů po transplantaci ledviny s dlouhodobě stabilní funkcí štěpu a skupina s chronickou rejekcí (úplný opak operační tolerance). Problémem, ale je, že pacienti z obou kontrolních skupin, byli léčeni imunosupresivní léky. Ty mohou ovlivnit expresi některých genů a představují tak matoucí faktory. Existují už statistické postupy kompenzace matoucích faktorů, ale je otázkou, zda jsou použitelné pro data ze sekvenování, kdy máme velký počet změřených transkriptů a malý počet vzorků (který je způsoben malým počtem pacientů s tímto onemocněním). V tomto případě pak dochází k překompensaci dat. Studentka proto vytvořila generátor umělých dat, na nichž se snažila zjistit, jak velký počet pacientů by byl dostatečný, aby k překompensaci nedocházelo. Na vygenerovaných datech pak ukázala, že pro překompensování dat je důležitý jak počet transkriptů (příznaků), tak počet pacientů ve skupině se stabilní funkcí. Pokud by nám stačilo měření jen 100 transkriptů, stačilo by nám 50 pacientů, při 10 000 transkriptech už by bylo potřeba 10 000 pacientů. Práce je metodicky dobře zpracovaná, škoda jen, že studentka ne zvolila pro sepsání angličtinu, protože stylisticky je plná chyb, což kazí celkový dojem.

Otázky:

1. Které z metod kompenzace matoucích faktorů, popsanych v úvodu (stratifikace, propensity score matching, regrese), jsou/nejsou vhodné pro náš konkrétní cíl.
2. Kapitola 3.7.3. Jak jste došla k tomu, že pravděpodobnost podávání imunosuprese je 60% (pokud jste klasifikovala 23 STA, 15 OT a CR, pak pravděpodobnost podávání IS je v celém souboru 28.8%)?
3. V kapitole 4.2. popisujete výpočet parametrů, potřebných pro klasifikátory pro 3 metody, ale ve výsledcích jste použila pouze metodu regrese, přičemž metoda SVM měla nejlepší AUC. Proč jste si vybrala regresi?



POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

4. Obrázek 4.6. Jestli správně chápu, porovnála jste skupinu OT vs STA+CR a hledala DEG s adj $p < 0.05$ a rozumnou $\log_2 FC$. A našla jste 81 diferenciólně exprimovaných transkriptů (předpokládáte, že mohou být ovlivněny IS, ale asi budou ovlivněny i samotnou třídou)

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **C - dobře**.

Datum: 11.1.2022

Podpis: