Posudek oponenta

autor posudku: Jan Stuchlý
autorka práce: Bc. et Bc. Nicole Aemilia Urban
Název práce: Interactive Clustering Approaches in Single-cell Cytometry

4. září 2024

Tématem předkládané práce je interaktivní clusterová analýza cytometrických dat a implementace příslušného interaktivního rozhraní. Jedná se o přepracovanou verzi práce neúspěšně obhajované v minulém roce.

Text práce je členěn do 3 kapitol. První kapitola popisuje principy průtokové cytometrie, analýzy cytometrických dat a výpočetních metod, které budou v práci využity (clustering, vizualizace a dimenzionální redukce cytometrických dat). Druhá kapitola popisuje implementaci nástroje „Ash“, který je hlavním výstupem práce. Třetí kapitola pak popisuje aplikaci interaktivního nástroje „Ash“ na reálná biologická data.

„Ash“ umožňuje analýzu předem provedeného hierarchického clusteringu a vizualizaci dat v podobě dendrogramu, heatmapy a projekcí clusterů na dimenzionální redukci či 2D scatterplot.

První kapitola je oproti minulé verzi dobře čitelná a základní principy jsou dobře pochopitelné. Text stále obsahuje některé nepřesnosti. Například autorka směšuje principy samotné průtokové cytometrie se sortováním („... the separation is achieved by passing the sample through narrow nozzle that vibrates.“ - tento proces se týká pouze sortování). Také nepovažuji za vhodné charakterizovat vlastnosti průtokové cytometrie na základě publikace čtvrt století staré („Usual throughput of a cytometer ranges between 100-1000 cells per second... Up to 10^5 of rows could be expected in a single dataset in resulting from one milliliter of the sample “). Dále viz otázky k obhajobě. Nicméně žádná z těchto nepřesností nebrání srozumitelnosti zbytku práce.

Kapitola 2 popisuje detaily implementace nástroje „Ash“ a porovnává „Ash“ jinými nástroji pro analýzu cytometrických dat. Přehled a porovnání knihoven vhodných pro tvorbu webových aplikací a vizualizaci dat byly, oproti minulé verzi, přesunuty do druhé kapitoly, kam logicky patří. Text byl zjednodušen je nyní daleko přehlednější. Pouze úvodní sekce „2.1 Overview of Interactive Clustering Tools for Flow Cytometry“ není příliš informativní - volba porovnávaných nástrojů se zdá být zcela náhodná - proč je posuzován nástroj, který není určen ke clusteringu (XCluSim) nebo zastaralý nástroj AUTOKLUS? Tato sekce by zasloužila větší pozornost.

Třetí kapitola popisuje aplikaci „Ash“ na reálná data. Problém je dobře zvolen a je demonstrováno, jak může heterogeneita v datech uniknout manuální analýze. Dále je ukázáno, jak implementovaná technika takovou heterogeneitu odhalí a umožní ji detailně popsat. Zároveň se ukazují slabiny implementace „Ash“, které ho činí obtížně použitelným pro reálná data. Data musela být rozdělena na části po 5000 buňkách a každá část analyzována zvlášť[1].

Celkové hodnocení: Celkově považuji práci za málo ambiciozní. „Ash“ není v tuto chvíli nástroj, který byl přímo použitelný na reálná data biology (nedokáže pracovat s dostatečně velkými daty, vlastní clustering je nutné provádět zvlášť a data pak exportovat). Zároveň práce nepřináší nové myšlenky a pouze demonstruje některé dobře známé principy analýzy cytometrických dat. Oceňuji dobře připravené postupy pro instalaci a deployment webové aplikace. Po formální stránce je současná verze zcela uspokojivá a přijatelná jako kvalifikační práce.

Otázky k obhajobě:

- V úvodu zmiňujete, že „Ash“ má oproti iDendro, kterým se inspirujete, lepší škálovatelnost na velká data - demonstrujte to prosím.

[1] Pokud bylo toto rozdělení provedeno pouze pro ilustraci, prosím autorku, aby při obhajobě demonstrovala, jak lze „Ash“ aplikovat na reálná data (tedy alespoň vyšší desítky tisíc eventů).

- Popište velmi stručně metody dimenzionální redukce, které v práci používáte a porovnejte je mezi sebou. V sekci 1.4 zmiňujete některé vlastnosti u jednotlivých metod - nicméně obtížnost interpretace nových proměnných ve vztahu k původním určitě není určující vlastnost PCA (spíše bych řekl, že hlavní komponenty lze interpretovat vcelku intuitivně). Pro t-SNE není specifické, že je to metoda neparametrická (ačkoli je to samozřejmě tak) a pro UMAP není specifické, že nekonverguje k lokálnímu minimu (lokálnímu minimu čeho přesně?).

Mgr. Jan Stuchlý, Ph.D.