Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Adam Šmelko, Bc.
Název práce	GPU-accelerated Mahalanobis-average hierarchical clustering
Rok odevzdání	2020
Studijní program	Informatika **Studijní obor** Softwarové a datové inženýrství

Autor posudku	Miroslav Kratochvíl **Role** vedoucí
Pracoviště	Katedra softwarového inženýrství

Text posudku:

Hlavním dosaženým cílem práce je zrychlení výpočtu komplikované varianty hierarchického clusterování pomocí GPU.

Akcelerace hierarchických clusterovacích algoritmů na GPU je obecně problematická, prakticky všechny dostupné algoritmy jsou navržené pro sekvenční výpočet. Práce se konkrétně zaměřuje na variantu s metrikou určenou Mahalanobiho vzdáleností, která je vhodná na relativně nízkodimenzionální data ($d < 30$) tvořící elipsoidní Gaussovské shluky. Výpočet této vzdálenosti mezi dvěma clustery závisí na výpočtu kovariance v jednotlivých clusterech a množství dalších maticových operací, což otevírá potenciál pro paralelizaci.

Hlavním pozitivním přínosem práce je implementace, která současnou, poměrně dobře optimalizovanou CPU implementaci (software `mhca`, algoritmus původně navrhli a implementovali Fišer a kol. (2012)) zrychluje o několik řádů; v závislosti na velikosti dat a způsobu clusterování autor naměřil zrychlení mezi 10× až 5000×. Výsledky jsou testované na realistických datech z hmotnostní cytometrie. Praktickým výsledkem práce je, že algoritmus jde použít na mnohem větší data než doposud, čímž pravděpodobně bude možné generovat přesnější biologické výsledky. Autor celkově prokázal, že dokáže samostatně analyzovat, sestavit a optimalizovat kvalitní software pro podporu jiných vědeckých odvětví.

Největším nedostatkem práce je nevyvážená úroveň detailu různých částí textu. Popisy použitých metod, implementace a hlavních výsledků práce (tj. zrychlení oproti původní implementaci algoritmu) je poměrně vyčerpávající; zajímavostí je vizuální srovnání dendrogramů pomocí stromové vzdálenosti jednotlivých datových bodů. Naproti tomu chybí diskuse některých očekavatelných témat:

- Přestože implementace očividně prošla několika vlnami optimalizací a měření — autor například ukazuje, že cachování většího množství nejbližších sousedů každého clusteru než 1 nemá s použitou implementační strategií smysl — text práce bohužel popisuje vliv velkého

množství zmiňovaných nízkoúrovňových optimalizací na celkovou rychlost výpočtu jen velice implicitně. Konkrétně je z autorova výstupu těžké určit, jak moc je rychlost ovlivněna layoutem dat a využitím atomických instrukcí, případně kolik času algoritmus běžně stráví zpracováním kterých částí výpočtu (tj. spouštěním kterých CUDA kernelů). Z textu práce zároveň není možné odvodit, jakou metodologii autor zvolil při optimalizaci, následkem čehož není možné ani tvrdit, jestli je implementace v nějakém kontextu optimální.

- Optimalizovaná verze algoritmu se od původní liší v 'malých detailech,' zběžně komentovaných mj. na straně 46 a v závěru práce. Tyto ale v případě testovacího datasetu 'Nillson rare' způsobily poměrně velký rozdíl ve výsledcích. To nemusí být nutně špatně, ale rozdíl by bylo užitečné komentovat detailněji, případně ukázat přímo na datech. Čtenář takto nemá možnost odvodit, jestli je rozdíl způsobený např. akumulací zaokrouhlovacích chyb, nějakou specifickou vlastností dat která má zásadní vliv na výpočet kovariance malých clusterů, nebo jde o artefakt existující jen v tomto jediném případě. Konkrétnější ukázky výsledků by byly vhodným doplňkem současného rozsahu práce.

Angličtina práce je na obvyklé úrovni, jazykové nedostatky jsou spíše syntaktické a stylistické než gramatické. Místy se vyskytují tvrzení, jejichž význam je i v kontextu těžké odhadnout, většinou kvůli nedostatečně specifikovaným větným členům — např. v popisu obrázku 1.1, ve větě 'The choice of a linkage criterion in hierarchical clustering algorithm is *vital*.' (strana 10), 'In the *means* of the proximity' (strana 12), nebo 'Last, a *special* transformation is performed on the inverse covariance matrix.' (strana 34). Tyto problémy by se spolu s několika drobnými faktickými nepřesnostmi daly odstranit důkladnější korekturou textu.

Přes zmíněné nedostatky předpokládám, že práce poskytne zajímavý základ pro další výzkum a publikace.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 24.6.2020

Podpis: