Posudek oponenta bakalářské práce

Autor/autorka práce: **Jakub Schenk**

Název práce: **Porovnání metrik pro shlukovací algoritmy**

Bakalářská práce se zabývá jedním shlukovacím algoritmem a to k-means. Pro tento algoritmus jsou v provedených testech používány tři různé metriky, tj. Euklidovská, Manhattanská a Geodetická.

V úvodu práce jsou popsány různé typy shlukovacích metod a příslušné metriky, které se ve shlukovacích algoritmech používají. Popis těchto kapitol je založen na šesti zdrojích, z nichž tři jsou pouze webové stránky. Práce s literaturou by zasloužila více citací v textu a ne pouze informaci o tom, že kapitola „Shlukování" je vypracována na základě pramenů [7-10].

V definici problému je uvedeno, že shlukování je cíleno pro geodetická data. Nikde v práci není ale uvedeno, jaké jsou důvody shlukování geodetických dat, jaké vlastnosti by měly výsledné shluky splňovat a jak se z výsledného výstupu shlukovacího algoritmu s danou metrikou určí, zda bylo dosaženo požadovaného výsledku.

V části experimentálních výsledků bylo provedeno mnoho testů pro různé vstupní data (umělá a reálná data). Nejprve bylo provedeno hodnocení funkčnosti jednotlivých metrik. V práci chybí porovnání, jaká metrika je pro jaký typ dat vhodnější nebo porovnání v čem je jaká metrika lepší. U každého výsledku je pouze zobrazeno jaký je výstup shlukování, ale chybí konkrétnější porovnání. Dále byla otestována funkčnost algoritmu k-means a také popsána a otestována možnost kombinace více metrik v určitém poměru.

Ve dvou následujících kapitolách (4.5 a 4.6) jsou popsány výsledky na reálných geodetických datech. Chybí zde bližší popsání v čem, se jaká metrika chová lépe a jaký výsledek shlukovacího algoritmu je právě ten vhodný pro další zpracování datového setu.

V poslední kapitole před závěrem jsou popsány časové náročnosti implementovaného algoritmu k-means s jednotlivými metrikami v závislosti na různých vstupních datech. Měření byly provedeny vždy pouze jednou pro daný datový set a danou metriku. Student sám uznává, že hodnoty jsou pouze orientační. Bylo by tedy vhodné měření provést vícekrát a do výsledných tabulek uvést vždy průměrné časy běhu algoritmu.

Zadání práce bylo splněno s menšími výhradami, kdy čtvrtý bod zadání („Dosažené výsledky popište v textu práce a zhodnoťte.") je sice splněn, ale očekával bych v práci podrobnější zhodnocení dosažených výsledků pro různé metriky použité na různých datových setech. Celkově je ale práce sepsána přehledně a během experimentů bylo provedeno velké množství testů. Zdrojový kód je psán čitelně a obsahuje potřebné komentáře pro pochopení jeho funkčnosti.

Dotazy k práci:

1. Jaké jsou důvody shlukování geodetických data a podle čeho určit, zda je výsledné shlukování „dobré“, nebo „špatné“?
2. Jaká metrika je podle vašich provedených experimentů vhodná pro jaký typ dat?
3. Bylo by možné použít kombinaci více metrik již během samotného výpočtu shlukovacího algoritmu? Ve vašem popisu jsou nejprve provedeny výpočty shluků pro každou metriku zvlášť a až tyto výsledky jsou kombinovány dohromady v určitém poměru.

Navrhuji hodnocení známkou **velmi dobře** a práci doporučuji k obhajobě.

V Plzni 28. 5. 2021 Ing. Michal Šmolík, Ph.D.