Oponentský posudek bakalářské práce

Název práce: Použití filtrovacích algoritmů ve shlukové analýze
Autor: Matěj Pacovský

Předložená práce se zabývá různými algoritmy shlukové analýzy pro klasifikaci vícerozměrných statistických dat. Jsou studovány algoritmy k-průměrů a x-průměrů a jejich modifikace využívající filtrovací stromy, metody jsou demonstrovány na reálných i umělých datech.

Práce je rozdělena do pěti kapitol. V první části autor shrnuje základy shlukové analýzy, ve druhé, třetí a čtvrté popisuje po řadě algoritmus k-průměrů, který třídí data iterativně do pevného počtu shluků, dále filtrovací algoritmus, pomocí kterého je možné výpočet urychlit na základě heuristiky a algoritmus x-průměrů, kde počet vzniklých shluků není fixován. Metody jsou vysvětleny přehledně a srozumitelně a demonstrovány na názorných příkladech. V páté kapitole je popsána samotná implementace algoritmů ve výpočetních programech a metody jsou vyzkoušeny na třech simulovaných datových souborech a na dvou příkladech z praxe. Výsledky získané pomocí jednotlivých algoritmů jsou vysvětleny a porovnány.

Celkově má práce velmi dobrou úroveň. Student látku nastudoval z více zdrojů, porozuměl ji a přehledně vysvětlil. Je třeba ocenit i softwarovou implementaci jednotlivých metod a příslušnou dokumentaci. Práce má několik nedostatků věcných i formálních, podstatná část je uvedená níže.

Seznam některých nedostatků a bodů k diskusi:

1. Některé základní pojmy jsou rozepsány i když by je stačilo zmínit (Eukleidovská vzdálenost), jiné zas vyloženy nejsou nebo chybí příslušná reference (Manhattanská metrika, Holmesův algoritmus).
2. Je škoda, že nebylo možné provést srovnání úspory reálného času běhu algoritmů. To je nahrazeno výpočtem očekávaného počtu operací.
3. Proč jsou v oddílu 5.2 při porovnávání algoritmů na datovém souboru A použity různé hodnoty parametru ϵ (v jednom případě 0.01 a ve dvou 0.5)?
4. V odstavci na str.34-35 se uvádí, že bylo nutné simulovat počáteční nastavení vícekrát, abychom dosáhli uspokojivého řešení - tento postup ale snižuje srovnatelnost výsledků.
5. Při srovnání výsledků pro různé hodnoty parametru *par* pro datový soubor B (str.35-37) vychází nejlépe $par = 100$, což je ale nejvyšší testovaná hodnota. Dá se usoudit, jaký by byl výsledek pro ještě vyšší hodnotu?
6. Z posledního odstavce závěru (str.47) není jasné, jak porovnání autorovy implementace algoritmu x-průměrů s postupem uvedeným v Pelleg a Moore (2000) souvisí s porovnáním s implementací programu Weka.
7. Tabulky (např. str.32,34,40) nejsou označeny, někdy také nemají popsané řádky a chybí v nich hodnoty, aniž by bylo vysvětleno proč a zda nejsou důležité.
8. Větší množství číslovaných výčtů (str.3,10,16,19,40) je rozděleno koncem stránky, což ubírá na přehlednosti, zvláště pak když se jedná o popis samotných algoritmů.
9. Některé grafy jsou nepřehledné (např. obr.5.8) a někdy chybí odkaz v textu (obr.5.7).
10. Seznam použité literatury nemá konzistenzní formát. Webové zdroje jsou bez popisu a data.

Domnívám se, že předložená práce splňuje nároky kladené na bakalářskou práci a doporučuji ji za ni uznat.

V Praze dne 10. června 2012 Mgr. Petr Novák