Centrála odběru krve - shlukování pomocí algoritmus k-means

Martin Svoboda

***Abstrakt*— Úkolem práce je naleznout nejlepšího nastavení parametrů pro shlukovací algoritmus k-means na datech z Centrály odběru krve na Taiwanu. Otestoval jsem kvalitu výsledků pro různá nastavení počtu centroidů a různé metodiky měření vzdálenosti. Došel jsem k závěru, že nejkvalitnější shluky vznikají při použití dvou centroidů a kosinové vzdálenosti.**

I. Zadání

Pro data získaná z Centrály odběru krve "Blood Transfusion Serice Center Data Set" nalézt nastavení parametrů pro k-means algoritmus. Nalezené nastavení algoritmu by mělo produkovat co nejlepší výsledky.

II. Úvod

Vstupní data obsahují záznamy o 748 dárců krve. Pro každého jsou k dispozici údaje o počtu měsíců od poslední návštěvy (Recent), celkovém počtu odběrů (Frequency), celkovém objemu darované krve (Monetary) a počtu měsíců od první návštěvy (Time). Mým úkolem je najít nastavení k-means algoritmu, tak aby produkoval co nejlepší výsledky.

III. Metodika

Pro zpracování dat jsem využil software Matlab 7.9 a připravené skripty ze sedmého cvičení předmětu Y336VD. Vstupní data jsem nejdříve normalizoval, aby hodnoty byly v rozmezí nula až jedna. Dále jsem z dat odstranil pátý sloupec, který udával kategorie.

Použil jsem skript kmeans4.m, který jsem upravil tak, aby vracel průměrnou hodnotu siluety pro různý počet centroidů. Navíc vrací hodnoty pří použití různých metrik měření vzdálenosti.

Dále jsem využil skriptu kmeans1.m, který vrací grafickou podobu siluety.

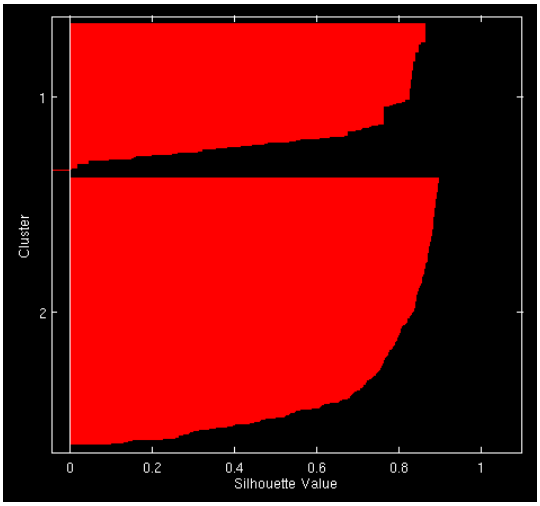
IV. Experimenty

Provedl měření určující nejlepší nastavení algoritmu k-means. Tabulka I ukazuje průměrné hodnoty siluety pro nastavení počtu centroidů od dvou do deseti. Dále ukazuje hodnoty při použití různých metodik meření vzdálenosti. Konkrétně se jedná o případy, kdy je použita euklidovská, manhattonská a kosinová vzdálenost. Všechny testy byly provedeny desetkrát s různým počátečním umístěním centroidů.

Obr. 1 ukazuje vykreslení konkrétní siluety, kdy počet centroidů je dva a pro měření vzdálenosti je použita kosinová vzdálenost.

TABULKA I
PRŮMĚRNÁ HODNOTA SILUETY PRO RŮZNÁ NASTAVENÍ ALGORITMU K-MEANS

	Metrika měření vzdálenosti		
Počet centroidů	Euklidovská	Manhattonská	Kosinová
2	0.6696	0.4182	0.7340
3	0.5297	0.3518	0.6042
4	0.5348	0.3323	0.5884
5	0.5139	0.3360	0.5667
6	0.5198	0.3400	0.6285
7	0.5161	0.3397	0.5915
8	0.5075	0.3162	0.5748
9	0.5120	0.3388	0.6274
10	0.4917	0.3379	0.6099



Obr. 1. Silueta pro k-means algoritmus. Počet centroidů je dva a pro měření vzdálenosti je použita kosinová vzdálenost.

V. Diskuze

Z prvního měření Tabulka I vyplynulo, že nejepší výsledky shlukování produkoval algoritmus k-means při nastavení počtu centroidů na dva a použití kosinové vzdálenosti. Z tohoto důvodu jsem vykreslil siluetu Obr. 1 právě pro toto nastavení. Je z ní vidět, že pro některá data je hodnota siluety v mínusu. Jedná se zřejmůě o odlehlé hodnoty, které by bylo dobré ze vstupních dat vymazat. Bohužel se mi nepodařilo odhalit, které konkrétní hodnoty jsou odlehlé, proto jsem je musel ve vstupních datech ponechat.

VI. ZÁVĚR

Podařilo se mi naleznout optimální nastavení pro algoritmus k-means. Za nejlepší parametry považuji nastavení počtu centroidů na dva a použití kosinové vzdálenosti jako metodiku pro měření odlehlosti vstupních hodnot. Toto nastavení vytváří shluky, jejichž průměrná hodnota siluety je 0.73. Průměrná hodnota siluety je snížena vlivem odlehlých hodnot, které jistě existují, ale nepodařilo se mi tyto hodnoty naleznout a vymazat.

REFERENCE

[1] I-Cheng Yeh: *Blood Transfusion Service Center Data Set*, Machine learning repository http://archive.ics.uci.edu/ml/, 2008
[2] Kordík P.: *LaTex šablona pro semestrální práci*, webové stránky předmětu Y336VD http://cw.felk.cvut.cz/doku.php/courses/y336vd
[3] Kessl R.: *K-means - skripty pro Matlab*, webové stránky předmětu Y336VD http://cw.felk.cvut.cz/doku.php/courses/y336vd