|  |
| --- |
| **TECHNICKÁ UNIVERZITA V KOŠICIACH****FAKULTA ELEKTROTECHNIKY A INFORMATIKY** |
| DokumentáciaPorovnanie metód zhlukovania: k-means vs. kohonenove sieteVýpočtová umelá inteligencie |
|  |  **Vypracovali:**  Martin Čertický Martin Kozák Jakub Tušan |

**Obsah**

[Zoznam obrázkov 3](#_Toc374991564)

[1 Teória 4](#_Toc374991565)

[2 Implementácia 7](#_Toc374991566)

[3 Porovnanie, performance 11](#_Toc374991567)

[3.1. Rýchlosti 11](#_Toc374991568)

[3.2. Úspešnosť zhlukovania 11](#_Toc374991569)

[4 Záver 13](#_Toc374991570)

Zoznam obrázkov

Obr. 1 - *k* počiatočných meansov (random) 5

Obr. 2 - vytvorených *k* zhlukov 5

Obr. 3 - geometrický stred zhlukov sa stáva novým meanom 5

Obr. 4 - opakujeme 2 a 3 až po dosiahnutie konvergencie 6

Obr. 5 - takto vyzeral výsledok pri 25 meansoch a iba 15-20 cykloch 7

Obr. 6 - ukážka dôležitosti správneho výberu počtu meansov (použitých iba 10 meansov) 8

Obr. 7 - rovnaký prípad ako pri Obr.7 - 10 000 bodov a iba 10 stredov 9

Obr. 8 - výsledný obrázok - 10 000 bodov a 20 meansov stačilo na uspokojivý výsledok 10

Obr. 9 - počet meansov nachádzajúcich sa mimo špirály po ukončení zhlukovania 12

1. Teória

„Algoritmus k-means vezme vstupný parameter k a rozdelí n objektov do k meansov tak, že maximalizuje podobnosť objektov v rámci meansu a minimalizuje podobnosť medzi meansami navzájom. V prvom rade, algoritmus náhodne zvolí k objektov, z ktorých každý reprezentuje mean - stred, centrum meansu. Ostatné objekty sú priradené ku meansom (reprezentovanými centrami meansov) na základe podobnosti, určenej cez vzdialenosti medzi objektmi a centrami meansov. Na základe rozdelenia objektov do meansov sa nanovo vypočítajú ich centrá (body rovnako vzdialené od objektov v meanse). Proces sa opakuje až kým nezkonverguje funkcia:

$$E= \sum\_{i=1}^{k} \sum\_{x\in C\_{i}}^{}|x-m\_{i}|^{2}$$

kde x je bod v priestore atribútov, reprezentujúci daný objekt.  mi sú centrá jednotlivých meansov Ci. Pri prepočte nových centier meansov sa neudeje zmena. Táto metóda pracuje dobre, keď sú meansy v dátach kompaktné zhluky, dobre navzájom oddelené. Tiež dobre pracuje na veľkých súboroch dát, pretože jeho výpočtová zložitosť je O(nkt), kde n je počet objektov, k počet meansov, t počet iterácií, obvykle platí k <<  n a t << n. „

Metóda má jednu vážnejšiu slabinu, tou je počiatočný náhodný výber meansov, čo v konečnom dôsledku nemusí znamenať termináciu pri rozdelení do najlepších meansov, kedy funkcia E nadobúda minimum. Z hľadiska rozdelenia do meansov, toto minimum môže byť len lokálne. Typickým príkladom je meansácia 4 bodov v rovine A,B,C,D, reprezentujúcich obdĺžnik so stranami *a, b b > a*. Body je potrebné meansovať na 2 meansy. Ako počiatočné centrá meansov sa určia stredy najdlhších strán *b*. Z toho vyplýva, že ide o situáciu, kedy funkcia *E*nadobudne minimum pri rozdelení do 2 meansov. Meansy sú reprezentované najdlhšími stranami obdĺžnika, keďže nový prepočet stredov nájde rovnaké stredy meansov.  Optimálne rozdelenie však predstavuje meansy reprezentované ako najkratšie strany obdĺžnika.

Metóda sa dá ale aplikovať len v prípade, keď vieme definovať stredy meansov. To býva problém, ak sa vo vzorkách vyskytujú nominálne atribúty. Metóda nie je vhodná na objavovanie meansov nekonvexných tvarov alebo veľmi rozdielnych veľkosti. Je tiež citlivá na znečistené alebo nekonzistentné dáta, pretože takéto vzorky značne ovplyvňujú výpočet centier meansov. Určenie vstupného parametra *k*, môže byť tiež problémom pri analyzovaní dát z domény, kde užívateľ nemá dostatok znalostí na predpokladanie počtu meansov, ktoré chce nájsť.



Obr. 1 - *k* počiatočných meansov (random)



Obr. 2 - vytvorených *k* zhlukov



Obr. 3 - geometrický stred zhlukov sa stáva novým meanom



Obr. 4 - opakujeme 2 a 3 až po dosiahnutie konvergencie

Teda ako sme už spomenuli vyššie, pri tejto metóde zhlukovania je veľmi dôležité správne určenie počtu meansov, z ktorých budeme vychádzať. Aby nedošlo k prípadu, že sa budú zbytočne zvyšovať nasilu počty iterácií pre prípady, kde nie je možné dosiahnuť dobré výsledky (kvôli malému počtu zvolených stredov).

1. Implementácia

Našou úlohou bolo rovnako ako pri prvom zadaní, vybrať ľubovoľný počet vzoriek z obrázku a pomocou zlukovania dosiahnuť aby sa nakoniec nachádzali vo vnútri špirály. Ukážeme si niekoľko obrázkov ako sme postupovali s našim programom:



Obr. 5 - takto vyzeral výsledok pri 25 meansoch a iba 15-20 cykloch

Testovali sme rôzne varianty (rôzny počet bodov na vstupe, rôzny počet meansov, počet cyklov), pričom sme si overili, že veľkú úlohu pri efektivite tohto zhlukovania tvorí počet stredov, ktoré na začiatku zvolíme. Je vhodné vedieť predpokladať nejakú približnú hodnotu vhodného počtu stredov namiesto použitia metódy pokus-omyl.



Obr. 6 - ukážka dôležitosti správneho výberu počtu meansov (použitých iba 10 meansov)



Obr. 7 - rovnaký prípad ako pri Obr.7 - 10 000 bodov a iba 10 stredov



Obr. 8 - výsledný obrázok - 10 000 bodov a 20 meansov stačilo na uspokojivý výsledok

1. Porovnanie, performance

## 3.1. Rýchlosti

Porovnávali sme metódy pri približne rovnakých číslach:

Čas, za ktorý sme schopní vytvoriť vyhovujúcu kohonenovu sieť pri 500 trénovacích bodoch: **1990ms** (400 cyklov, sieť veľká 5x5)

Časy a počty cyklov potrebné pre dobré zozhlukovanie pomocou k-means:

|  |  |  |
| --- | --- | --- |
|  | Počet cyklov | Čas(ms) |
|  | 16 | 104 |
|  | 12 | 93 |
|  | 13 | 104 |
|  | 17 | 112 |
|  | 13 | 101 |
|  | 15 | 103 |
|  | 15 | 108 |
|  | 25 | 133 |
|  | 10 | 96 |
|  | 9 | 91 |
| Medián | 14 | **103,5** |

## 3.2. Úspešnosť zhlukovania

Uvažovali sme navyše počet meansov, ktoré sa po skončení zhlukovania nachádzali na nežiadúcich pozíciách (mimo špirály), pričom výstupom sú priemerné hodnoty z 20 pokusov. Na Obr.9 je vidieť, že najlepšie hodnoty (najmenej „vysypaných“ meansov) sme dostali pri počte približne 20-25 meansov. Niekoľko vybraných hodnôt je vidieť aj v tabuľke:

|  |  |  |
| --- | --- | --- |
| Všetky meansy | Meansy mimo |  % |
| 2 | 1,8 | 90 |
| 10 | 8,4 | 84 |
| 15 | 6,2 | 41,33 |
| 20 | 3,8 | 19 |
| 24 | 3,3 | 13,75 |
| 25 | 3,85 | 15,4 |
| 30 | 3,4 | 11,33 |
| 40 | 4,65 | 11,63 |
| 50 | 7,26 | 14,52 |
| 100 | 26,75 | 26,75 |

****

Obr. 9 - počet meansov nachádzajúcich sa mimo špirály po ukončení zhlukovania

1. Záver

Po porovnaní obidvoch zhlukovacích metód môžeme v niekoľkých vetách zhrnúť:

Pri metóde k-means dostávame graf úspešnosti podobný parabole (keď prestrelíme hodnotu meansov, začne sa zhoršovať), kde najnižsia hodnota grafu predstavuje optimálne *k*. Teda počet meansov na zhlukovanie pre danú množinu dát.

 Pri použití kohonenovej siete je to s neúspešnými vzorkami naopak, keďže s narastajúcim počtom uzlov nám klesá počet vzoriek mimo žiadané hodnoty. Teda má zmysel zvyšovať počet uzlov až kým sa všetky body nenachádzajú na požadovaných pozíciách (na špirále).