**Hledání podskupin v datech**

**Milan Poláček**

# Zadání

## Cíl

Cílem druhého úkolu je seznámit se s praktickým použitím shlukovacích metod. Shlukovací metody lze využít k odhalení vnitřní struktury dat. Poznání vnitřní struktury dat může pomoci k lepšímu porozumění datům a jevům, které zachycují, stejně tak ke konstrukci lepších klasifikátorů.

## Data

Data byla nasbírána v roce 1993 a pokrývají charakteristiky a cenu vybraných 93 osobních aut. U každého auta byla zjišťována cena, spotřeba, výkon, hmotnost, rozměry a další charakteristiky (viz Tabulka příznaků z úplného zadání). Kompletní data jsou k dispozici v souboru cars.csv.

## Hledání podskupin v datech

Pokuste se v datech nalézt vnitřní strukturu (podskupiny) na základě hodnot příznaků. Zvolte vhodnou metodu. Zamyslete se nad tím, zda je vhodné data normalizovat. Výsledky vizualizujte a pokuste se je interpretovat, např. pomocí typických (průměrných) reprezentantů. Prozkoumejte, které z příznaků se nejvíce podílí na rozdělení instancí do jednotlivých shluků. Uvažte, jestli je možné dát nalezenou strukturu do souvislosti s klasifikací výrobce aut.

## Požadované kroky analýzy [10 bodů]

• upravte data pro shlukování [1 b]

• zvolte vhodnou metodu shlukování [1 b]

• zvolte vhodné parametry shlukovací metody [1 b]

• prezentujte výsledky shlukování pomocí typických reprezentantů shluků [3 b]

• zjistěte v jakých příznacích a jak se jednotlivé shluky liší (které příznaky jsou relevantní pro

hledání výrobce aut) [2 b]

• porovnejte výsledky shlukování s klasifikací výrobce aut [2 b]

# Úvod

K problematice shlukování z větších dat do podskupin jsem se věnoval již na bakalářském studiu, ale data vždy obsahovala maximálně 4 příznaky a 4 skupiny, pro které stačilo využít rozdělování do podskupin pomocí k-nn.

Proto jsem zde využil literatury dostupné na internetu, abych se s danou problematikou lépe seznámil.

# Metody

V této práci se dalo rozhodnou mezi dvěma hlavními klasifikátory. A to metodou Hierarchických shluků a metodou K-means. Jelikož metoda K-means vyžaduje znalost počtů podskupin zvolil jsem proto Hierarchické hluky (funkce *hclust*). Použil jsem dvě metody pro funkci *hclust*. Jednalo se o metodu *ward.D*, která hledá souvislé sférické (kulaté) shluky. Druhá metoda, kterou jsem zvolil, byla *complete.* Tato metoda se snaží najít a propojovat podobné shluky.

Nejprve jsem data musel zbavit záznamů, kde chyběli příznaky. Po jejich odstranění nám zbylo 82 záznamů.

Pro odhady dendrogramů jsem vyloučil příznaky Manufacturer, Model které jsou logicky unikátní. Dále jsem vyloučil příznaky Min.Price a Max.Price. Tyto dva příznaky podle pokusů zásadně neovlivňovali rozdělování dat do shluků. Vyloučení jiných jednoho nebo více příznaků většinou zapříčinilo separaci jednoho až dvou vozidel jako samotnou skupinu.

Data bylo dále nutno normalizovat, a proto jsem využil metodu *scale.* Pro zpřesnění výsledku jsem pro normalizaci využil odečtení mediánu a dělení střední průměrnou odchylkou.

Poté už lze spočítat matici vzdáleností, kde jsem využil základní funkci *dist*. Ačkoliv v knihovnách R existuje funkce *daisy,* rozhodl jsem se pro jednoduší funkci *dist* a to se základním nastavením (eukleidovská atd.). Zde bylo také nutno vzdálenosti označené jako *Inf* (nekonečno) nahradit jiným vysokým číslem (konkrétně 100), aby bylo možno matici vzdáleností využít pro funkci *hclust*.

Pro dělení dat pro výpočet typických reprezentantů jsem využil funkci *cutree*.

# Výsledky

Jak je vidět z dendrogramů, každá metoda vytvořila jiný počet shluků. Metoda *ward.D* (viz obrázek 2) do dvou hlavních shluků. Oproti tomu metoda *complete* (viz obrázek 1) je rozdělila do tří hlavních shluků. Pro výpočet typických průměrných reprezentantů jsem se rozhodl rozdělit data vždy rozdělit do n+1 shluků pomocí *cutree*. Kdy n reprezentuje počet viditelných hlavních podskupin v dendrogramech.

V tabulkách 1 a 2 je vidět reprezentantce vybraných příznaků reprezentantů.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Shluk | Freq | Price | MPG.city | MPG.highway | DriveTrain | AirBags | Cylinders | EngineSize | HP | Origin |
| 1 | 25 | 10,1 | 29 | 33 | Front | D&P | 2 | 1,6 | 92 | Non-USA |
| 2 | 23 | 18,2 | 21 | 28 | Front | D | 2 | 2,5 | 150 | Non-USA |
| 3 | 19 | 17,7 | 21 | 28 | Front | D | 2 | 2,3 | 135 | USA |
| 4 | 15 | 30 | 18 | 26 | Rear | D | 4 | 3,8 | 190 | USA |

Tabulka 3. Rozdělení shluků podle complete

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Shluk | Freq | Price | MPG.city | MPG.highway | DriveTrain | AirBags | Cylinders | EngineSize | HP | Origin |
| 1 | 22 | 10 | 29 | 33,5 | Front | D&P | 2 | 1,55 | 90 | Non-USA |
| 2 | 36 | 24,05 | 19 | 26 | Front | D | 4 | 3,35 | 170 | USA |
| 3 | 24 | 15,7 | 23 | 30 | Front | D | 2 | 2,2 | 119 | USA |

Tabulka 2. Rozdělení shluků podle ward.D

|  |
| --- |
| C:\Users\Milhouse\Documents\Scholla\___FEL\21rocnik\DVZ\AutaRB\img shluk\Complete Hclust.png |
| Obrázek 1. Dendrogram metody complete |

|  |
| --- |
| C:\Users\Milhouse\Documents\Scholla\___FEL\21rocnik\DVZ\AutaRB\img shluk\Ward Hclust.png |
| Obrázek 1. Dendrogram metody ward.D |

# Závěr

Podle výsledků nelze s určitostí říci, zda byli nesprávně použity metody nebo byla data nedostatečně upravena pro shlukování anebo to ze zadaných dat pomocí nejde jednoznačně určit. Shluky (podskupiny) podle typických průměrných reprezentantů totiž vykazují velkou rozmanitost a nelze tak s určitostí říci který příznaky jsou vhodné pro specifické hledání podskupiny (např. hledání výrobce aut, nebo země původu).

Ale lze například při zkoumání dendrogramů zjistit, že některé shluky pomocí metody *complete* mají společné prvky jako např. podobný typ auta (nižší střední třída, vyšší třída), země původu atd.

Protože jsem v problematice dělení aut do tříd znalý, vím že nejedna automobilka se snaží atraktivitu auta navýšit výbavou, která je typická pro jinou třídu (skupinu) aut. A proto se dá předpokládat, že klíč pro správné třízení automobilů vlastně neexistuje.