StatSoft

Shlukování podobných v softwaru *STATISTICA*

Tímto článkem nakoukneme do oblasti statistiky zabývající se shlukováním. Tedy situací, kdy chcete data/objekty nějak seskupit na základě jejich podobnosti. Článek je doplněn praktickým příkladem, který Vám ukáže, jak je to jednoduché.

Shluková analýza

Jedná se o metodu, která shlukuje objekty do skupin tak, aby objekty náležící do jedné skupiny, byly sobě podobné.

Praxe

Úloh a situací, kdy potřebujete objekty shlukovat do skupin, je nepřeberné množství. Abychom ukázali, co všechno se dá dělat, uveďme několik praktických využití:

Shlukování genů s podobnými vlastnostmi exprese.

Tvorba skupin studentů se stejnými vlastnostmi.

Shlukování otázek v dotazníku, na které respondenti odpovídají podobně.

Shlukování chemických prvků, které se chovají v nějaké situaci podobně.

Shlukování oblastí například podle spáchaných trestných činů.

Shluková analýza je často využívaná v marketingu. Například shlukování zákazníků do skupin na základě dat z dotazníkových šetření.

Shlukování uživatelů sociálních sítí může odhalit komunity lidí.

Shlukování produktů podle jejich vlastností.

…

Hierarchické shlukování

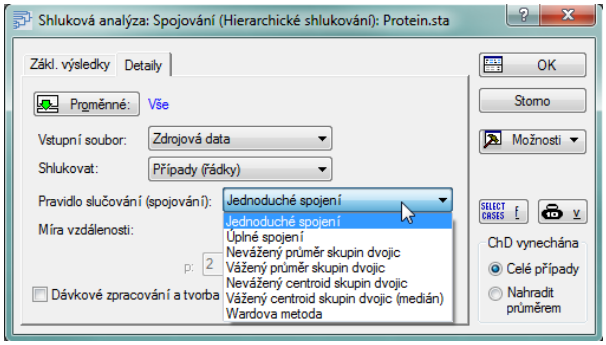
V našem článku a příkladu se zaměříme jen na jednu z metod shlukování a to na hierarchické shlukování.

Hierarchické shlukování znamená, že shluky vytváříme postupně v několika krocích. Na začátku máme n shluků (shluky s jedním prvkem). Ve druhém kroku máme $n-1$ shluků, protože jeden shluk sloučíme s nějakým jiným. Shluky, které se spojily, jsou ty, které mají mezi sebou nejmenší vzdálenost. V dalších krocích postupujeme analogicky až do vytvoření jednoho velkého shluku, který obsahuje všechny objekty (takovémuto hierarchickému shlukování se říká aglomerativní - objekty se postupně slučují). Rozvrh shlukování se v těchto modelech vyjadřuje nejčastěji pomocí grafického zobrazení nazývaného dendrogram (bude ukázán a dovysvětlen níže). Abychom toto shlukování mohli provést, potřebujeme si nadefinovat, jakou vzdálenost budeme používat a také odkud ve shluku se bude měřit vzdálenost k jinému shluku (případně jak se bude měřit vzdálenost mezi shluky). Začněme s tímto.

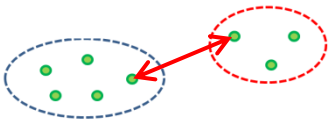
Pokud se podíváte na obrázek vlevo a zkusíte se zamyslet nad tím, co by mohla být vzdálenost mezi těmito dvěma shluky, zjistíte, že to není zase tak jednoduché, jak by se na první pohled mohlo zdát. Jistě Vás napadne mnoho možností, mezi kterými místy vzdálenost měřit.

Trocha teorie - Vzdálenosti mezi shluky

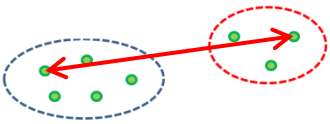
Výčet možností definování vzdálenosti mezi shluky bychom provedli popořadě, jak je to nastaveno v softwaru *STATISTICA*. Pokud si otevřete dialog shlukové analýzy (***Statistiky-Vícerozměrné statistiky-Shluková analýza-Spojování (hierarchické shlukování)***) záložku ***detaily***, nalezneme rozevírací seznam u položky *Pravidla slučování:*



Jednoduché spojení (single linkage) - vzdáleností dvou shluků je vzdálenost dvou nejbližších objektů z různých shluků.

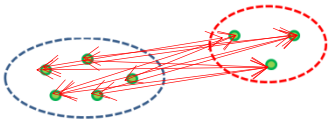


Úplné spojení (complete linkage) – toto je opačný extrém, zde vezmeme vzdálenost dvou nejvzdálenějších objektů.



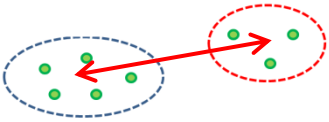
Neváženýprůměr skupin dvojic (UPGMA) – vzdálenost dvou shluků je průměrem vzdáleností všech dvojic (každý člen dvojice je z jiného shluku).

Vážený průměr skupin dvojic (WPGMA) – stejné jako výše, jen se jedná o vážený průměr vzdáleností všech dvojic – je brána v potaz velikost shluků. Vážené metody doporučovány v případě, že se dají očekávat rozdílné velikosti shluků.



Neváženýcentroid skupin dvojic (UPGMC) – vzdálenost dvou shluků je vzdáleností mezi centroidy shluků.

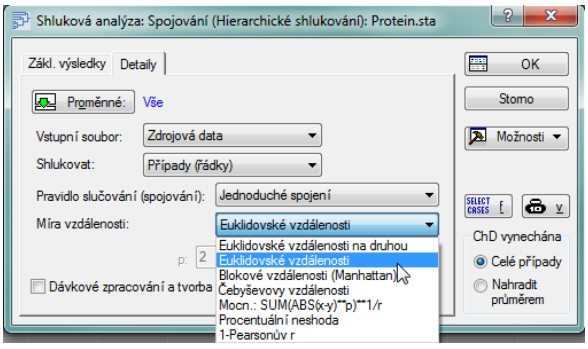
Vážený centroid skupin dvojic (WPGMC) – vážená vzdálenost dvou centroidů (váhy se určují podle velikosti shluků).



Wardova metoda – odlišný přístup oproti předešlým, založený na principu analýzy rozptylu. Počítá součet druhých mocnin odchylek případů v potenciálním sloučeném shluku od centroidu. Sloučí ty dva shluky, pro které je součet nejmenší.

Trocha teorie – Typy vzdáleností

Bavili jsme se o tom, jak určovat, jak měřit vzdálenost mezi shluky. Nyní se podíváme na typy vzdáleností samotných. Při vybírání typu vzdálenosti budeme mít v praxi situaci jednodušší než při určování, jak měřit vzdálenost mezi shluky, poněvadž podle povahy dat bychom typ vzdálenosti měli zvolit celkem jednoznačně.



Euklidovské vzdálenosti, Euklidovské vzdálenosti na druhou – jednoduše vzdálenost mezi body se spojitými hodnotami, klasická vzdálenost bodu od bodu, jak ji známe.

Bloková vzdálenost – vzdálenost, jako bychom se pohybovali po pouze vodorovně a svisle, ne šikmo (někdy se jí také říká Manhattanská vzdálenost).

Čebyševova vzdálenost – maximální rozdíl souřadnic v jednom rozměru.

Procentuální neshoda – podíl shodných prvků (souřadnic) mezi objekty a počtu všech prvků (tedy dimenze objektu). Používá se pro diskrétní data.

Ostatní možnosti jsou jasné již z názvu možnosti. Pokud byste potřebovali výpočetní detaily, odkážeme Vás na popis v nápovědě softwaru, najdete jej v sekci: ***Joining (Tree Clustering) Introductory Overview - Distance Measures***

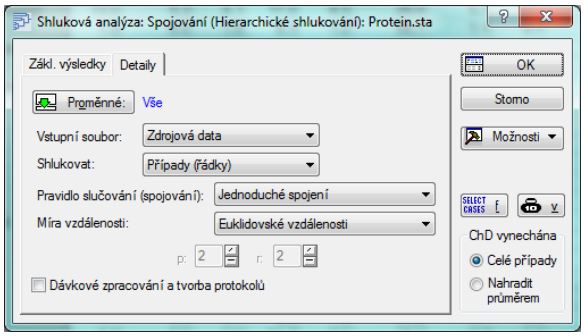
Příklad

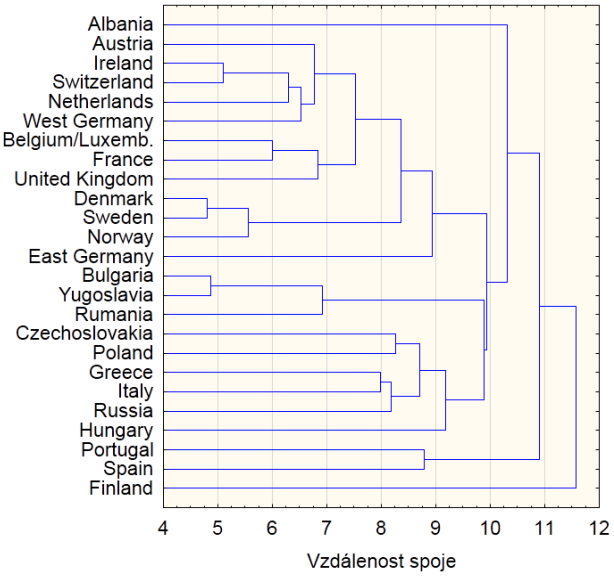
Máme k dispozici data *Protein.sta* (ta najdete v příkladech softwaru *STATISTICA*: ***Data-Otevřít příklady-Datasets***). Data představují odhadnutý příjem proteinů z 9 zdrojů obyvatelů ve 25 zemích Evropy. Data jsou z roku 1973. U těchto dat se můžeme ptát, které státy jsou v souvislosti s rozložením proteinů ve stravě podobné. Toto je tedy naše úloha, na kterou se přímo vybízí použití shlukové analýzy.

	1 MEAT	2 PIGPOUL	3 EGGS	4 MILK	5 FISH	6 CEREALS	7 STARCH	8 NUTS	9 FRUITVEG
Albania	10,1	1,4	0,5	8,9	0,2	42,3	0,6	5,5	1,7
Austria	8,9	14,0	4,3	19,9	2,1	28,0	3,6	1,3	4,3
Belgium/Luxemb.	13,5	9,3	4,1	17,5	4,5	26,6	5,7	2,1	4,0
Bulgaria	7,8	6,0	1,6	8,3	1,2	56,7	1,1	3,7	4,2
Czechoslovakia	9,7	11,4	2,8	12,5	2,0	34,3	5,0	1,1	4,0
Denmark	10,6	10,8	3,7	25,0	9,9	21,9	4,8	0,7	2,4

Uvidíte, že použití metody je naprosto jednoduché.

Spustíme ***Statistiky-Vícerozměrné statistiky-Shluková analýza-Spojování (hierarchické shlukování)-OK*** . Vybereme všechny proměnné, dále chceme shlukovat případy (státy) – obecně je možné si vybrat, jestli chceme shlukovat případy nebo proměnné (shlukování proměnných má jistě v některých příkladech svůj prokazatelný smysl). Vzdálenost necháme Euklidovskou a pravidlo slučování necháme pro začátek také tak, jak bylo přednastaveno. Klikneme na ***OK***. Nyní si již můžeme vygenerovat dendrogram.



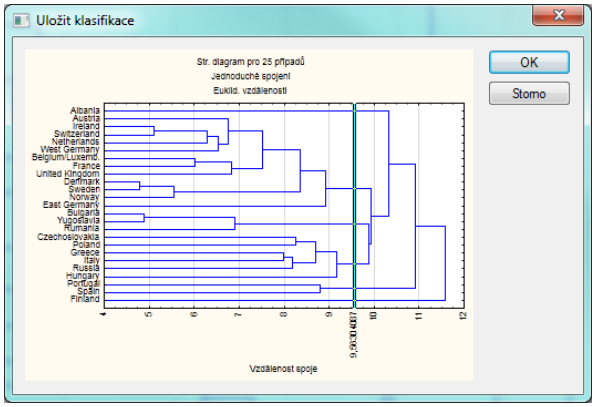


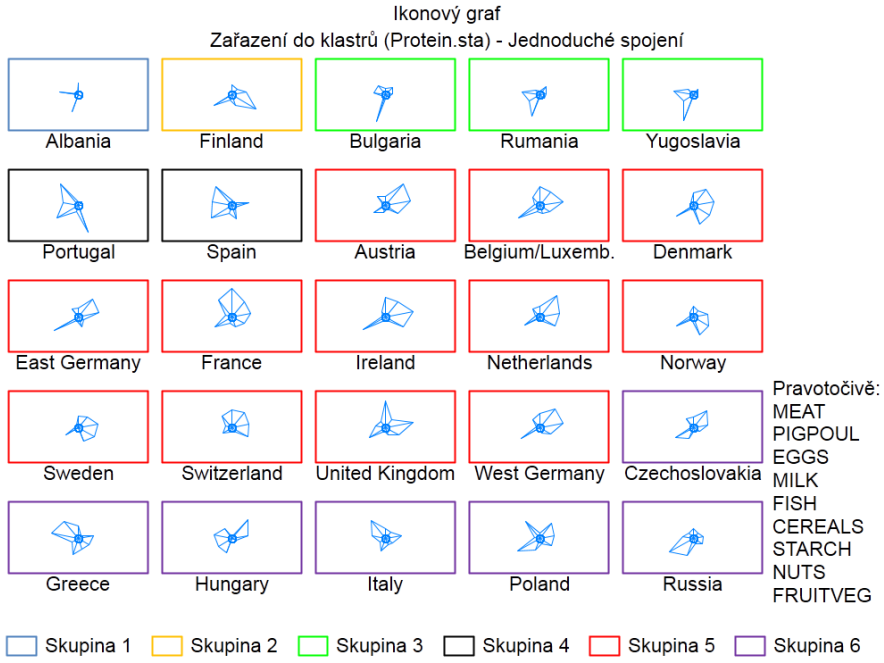
Vysvětlení dendrogramu – tento graf ukazuje kompletní historii spojování do shluků od jednotlivých objektů (vlevo) až to jednoho shluku se všemi objekty (vpravo). Na ose x je vzdálenost, při které se shluky spojily. Vidíme tedy, že první se spojily státy Dánsko a Švédsko, to se dá interpretovat tak, že mají tyto státy velmi podobné rozložení proteinů ve stravě. Spojily se ve vzdálenosti 4,8. Jako poslední se s ostatními spojilo Finsko, z čehož můžeme usuzovat, že je nejdále od ostatních.

Takovýto graf řekne hodně, nicméně to většinou není finální výsledek, finálním výsledkem je většinou rozdělení na několik shluků. K tomu nám poslouží **možnost** *Uložit klasifikaci* na záložce *Detaily*. Objeví se nám nový graf s posuvnou linií, která určuje místo dělení (shluky, které byly spojeny před touto čarou, budou patřit do jednoho shluku). Naše dělení rozdělilo soubor na 6 shluků. Vidíme, že Albánie a Finsko jsou samostatně, v klastru číslo 3 jsou Jugoslávie, Rumunsko a Bulharsko, u této skupiny vidíme oproti

	Zařazení do klastrů
Albania	1
Finland	2
Bulgaria	3
Rumania	3
Yugoslavia	3
Portugal	4
Spain	4
Austria	5
Belgium/Luxemb.	5
Denmark	5
East Germany	5
France	5
Ireland	5
Netherlands	5
Norway	5
Sweden	5
Switzerland	5
United Kingdom	5
West Germany	5
Czechoslovakia	6
Greece	6
Hungary	6
Italy	6
Poland	6
Russia	6

ostatním velký přísun proteinů z cereálií a velmi malý z ryb – toto mohou být proměnné, které předurčují tyto státy být daleko od ostatních. Čtvrtá skupina jsou Portugalsko a Španělsko – blízké země s podobnými stravovacími návyky. Další skupina je v zásadě západ a sever Evropy, skupina obsahující i Československo zase střed, východ a jih Evropy.





A to je v zásadě vše. Státy jsme rozdělili do skupin a stačilo opravdu jen několik málo kliknutí myší. Navíc rozdělené skupiny, zdá se, dávají smysl. Podívejme se ještě na ikonový graf všech proměnných s kategorickým odlišením skupin v rámečcích. Použili jsme ikonový graf typu *Hvězdy* s označením ikon na základě výsledků shlukování (obrázek vlevo). Ikonové grafy jsou vhodné pro zobrazení vícerozměrných dat (viz **článek o Chernoffových tvářích**), z grafu je vidět, že Albánie má naprosto rozdílné chování než ostatní státy, proto je také identifikována jako samostatný shluk. Podobně Finsko je „zvláštní" ve směru „mléka", atd.

Poznámka: My jsme spojovali shluky na základě jednoduchého spojování. Pokud není nějaký konkrétní důvod, proč použít právě tuto metodu, tak se doporučuje zkusit vytvořit shluky podle více metod. Pokud se struktura shluků pro různé metody opakuje, pak shlukování zachytilo strukturu dat správně. Pro kontrolu síly (kvality) shlukování se tedy doporučuje vyzkoušení více typů shlukování.

V tomto příkladu další zkoušení necháme již na Vás, podle návodu výše by to pro Vás neměl být žádný problém.

Závěr

Ukázali jsme si, jak provést základní výpočet, co znamenají nastavení metody hierarchického shlukování v softwaru, jak číst výsledky, sepsali jsme Vám několik rad. Nicméně je potřeba upozornit, že toto je pouze úvodní článek, který Vás měl zasvětit to této tématiky a který obsahuje pouze základní teorii k tomuto tématu. Doufáme, že článek ve Vás probudil chuť shlukovou analýzu využívat. Pokud byste se chtěli dozvědět více, můžete například navštívit náš kurz **Vícerozměrných statistických metod**.

StatSoft STATISTICA

StatSoft CR s.r.o.
Ringhofferova 115/1, 155 21 Praha 5 – Zličín
t +420 233 325 006, **f** +420 233 324 005
e info@statsoft.cz