4.6.2 Analýza shluků CLU

Analýza shluků (**Clu**ster analysis, CLU) patří mezi metody, které se zabývají vyšetřo-váním podobnosti *vícerozměrných objektů* (tj. objektů, u nichž je změřeno větší množství proměnných) a jejich klasifikací do tříd čili *shluků.* Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. Podle způsobu shlukování se postupy dělí na *hierarchické* a *nehierarchické*. Hierarchické se dělí dále na *aglomerativní* a *divizní.*

Hierarchické postupy jsou založeny na postupném spojování objektů a jejich shluků do dalších, větších shluků. Nejprve se vypočte základní matice vzdáleností mezi objekty. U *aglomerativního shlukování* se dva objekty, jejichž vzdálenost je nejmenší, spojí do prvního shluku a vypočte se nová matice vzdáleností, v níž jsou vynechány objekty z prvního shluku a naopak tento shluk je zařazen jako celek. Celý postup se opakuje tak dlouho, dokud všechny objekty netvoří jeden velký shluk nebo dokud nezůstane určitý, předem zadaný počet shluků. *Divizní postup* je obrácený. Vychází se z množiny všech objektů jako jediného shluku a jeho postupným dělením získáme systém shluků, až skončíme ve stadiu jednotlivých objektů. Výhodou hierarchických metod je nepotřebnost informace o optimálním počtu shluků v procesu shlukování; tento počet se určuje až dodatečně. Při shlukování vznikají dva základní problémy:

(a) *způsob měření vzdáleností mezi objekty.* I když existuje celá řada měr vzdáleností (vícerozměrných metrik), nejčastěji se užívá *euklidovská metrika*, která je přirozeným zobecněním běžného pojmu vzdálenosti;

(b) *volba vhodné shlukovací procedury* dle zvoleného způsobu metriky.

Metody metriky shlukování jsou

Metoda průměrová (Average): vzdálenost dvou shluků se počítá jako průměr z možných mezishlukových vzdáleností dvou objektů, kdy se mezishlukovou vzdáleností objektů rozumí vzdálenost dvou objektů, z nichž každý patří do jiného shluku. Nejbližší jsou shluky, které mají nejmenší průměrnou vzdálenost mezi všemi objekty jednoho a všemi objekty druhého shluku. Dendrogramy mají strukturu podobnou dendrogramům metody nejvzdálenějšího souseda, pouze spojení je provedeno při obvykle vyšších vzdálenostech.

Metoda centroidní (Centroid): vzdálenost shluků se počítá jako euklidovská vzdálenost jejich těžišť. Nejbližší jsou ty shluky, které mají nejmenší vzdálenost mezi těžišti.

Metoda nejbližšího souseda (Single, Nearest): kritériem pro vytváření shluků je minimum z možných mezishlukových vzdáleností objektů. Metoda tvoří nový shluk na základě nejkratší vzdálenosti mezi shluky (či objekty) a neumí proto rozlišit špatně separované shluky. Je zde silná tendence ke tvorbě řetězců. Jsou-li objekty na opačních koncích řetězce zcela nepodobné, řetězování může vést až ke zcela mylným závěrům. Na druhé straně je to jedna z mála metod, která umí roztřídit a rozlišit i neeliptické shluky.

Metoda nejvzdálenějšího souseda (Complete, Furthest): počítá vzdálenost dvou shluků jako maximum z možných mezishlukových vzdáleností objektů. Probíhá podobně jako metoda Single s jednou důležitou výjimkou: vzdálenost (či nepodob-nost) mezi shluky je určována vzdáleností (či nepodobností) mezi dvěma nejvzdá-lenějšími objekty, každý přitom je z jiného shluku. Proto všechny objekty ve shluku jsou klasifikovány na základě maximální vzdálenosti či minimální podobnosti vůči objektům ve druhém shluku.

Metoda mediánová (Median): jde o jisté vylepšení centroidní metody, neboť se snaží odstranit rozdílné "váhy", které centroidní metoda dává různě velkým shlukům.

Wardova metoda je založena na minimalizaci ztráty informace při spojení dvou tříd. V každém kroku je uvažován takový možný pár objektů (či shluků), aby suma čtverců odchylek od střední hodnoty $ESS = \sum_{i=1}^{n} (x_i - \bar{x})^2$ dosáhla při vzniku shluku svého minima.

Nehierarchické shlukovací metody: u *metody typických bodů* (Seeded) uživatel na základě svých věcných znalostí určí, které objekty mají být "typickými" představiteli nově vytvořených shluků a systém rozdělí objekty do shluků podle jejich euklidovské vzdálenosti od těchto typických objektů. V nehierarchických shlukovacích metodách je počet shluků obvykle předem dán, i když se v průběhu výpočtu může změnit. Zůstává-li počet shluků zachován, hovoříme o nehierarchických metodách s *konstantním počtem shluků*, v opačném případě o nehierarchických metodách s *optimalizovaným počtem shluků*. Nehierarchické metody zahrnují dvě základní varianty - optimalizační metody a analýzu módů, medoidů. *Optimalizační nehierarchické metody* hledají optimální rozklad přeřazováním objektů ze shluku do shluku s cílem minimalizovat nebo maximalizovat nějakou charakteristiku rozkladu. Metody, označované jako *analýza módů, medoidů*, představují hledání rozkladu do shluků, kde shluky jsou chápány jako místa se zvýšenou koncentrací objektů v *m*-rozměrném prostoru proměnných.

Místo výchozí matice vzdáleností může být v některých případech ke shlukování použita i *korelační matice.*

(a) Hierarchické shlukování

Analýza shluků patří mezi metody, zabývající se vyšetřováním podobnosti vícerozměrných objektů (tj. objektů, u nichž je změřeno větší množství proměnných) a jejich roztříděním do tříd čili *shluků.* Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. Analýzou shluků budeme sledovat a vyšetřovat jednak podobnost objektů, analyzovanou pomocí *dendrogramu objektů*, a jednak podobnost proměnných analyzovanou pomocí *dendrogramu proměnných.*

Dendrogram, diagram shluků nebo vývojový strom se objeví pouze v případě zadání hodnot původních proměnných a nikoli při zadání maticí vzdáleností. Výsledkem je zobrazení hodnot ve dvojrozměrném prostoru, kde osy tvoří zadané proměnné. Objeví se také “obkroužení” objektů v jednotlivých shlucích.

Dendrogram podobnosti objektů je standardní výstup hierarchických shlukovacích metod, ze kterého je patrná struktura objektů ve shlucích.

Dendrogram podobnosti proměnných odhaluje nejčastěji dvojice či trojice (obecně *m*-tice) proměnných, které jsou si velmi podobné a silně spolu korelují. Odhaluje proměnné, které jsou ve společném shluku, které jsou si tím pádem značně podobné a které jsou také vzájemně nahraditelné. To má značný význam při plánování experimentu a respektování úsporných ekonomických kritérií. Některé vlastnosti (či proměnné) není třeba vůbec měřit, protože jsou snadno nahraditelné jinými a nepřispívají do celku velkou vypovídací schopností.

Míra věrohodnosti: dendrogram lze sestrojit celou řadou technik. Prvním kritériem věrohodnosti čili těsnosti proložení při volbě “nejlepšího dendrogramu”, jež nejlépe odpovídá struktuře objektů a proměnných mezi objekty, je *kofenetický korelační koeficient CC*. Je to Pearsonův korelační koeficient mezi skutečnou a predikovanou vzdáleností, založenou na dendrogramu. Je-li tato hodnota větší než 0.75, je obvykle nulová hypotéza o dané struktuře zamítnuta. Hodnota 0.9 svědčí, že dendrogram vůbec neodpovídá skutečné struktuře dat.

Druhým kritériem těsnosti proložení je *kritérium delta* Δ, které měří stupeň přetvoření, distorze spíše než stupeň podobnosti. Kritérium delta je definováno vztahem

$$\Delta_A = \left[\frac{\sum_{j<k}^{N} |d_{jk} - d_{jk}^{*}|^{1/A}}{\sum_{j<k}^{N} (d_{jk}^{*})^{1/A}} \right]^{A},$$

kde $A = 0.5$ nebo 1 a d^{*}_{ij} je vzdálenost získaná z dendrogramu. Jsou žádoucí hodnoty *delta* blízké nule. Řada autorů ukázala, že metoda průměrová vede obvykle k nejlepšímu dendrogramu.

Postup shlukové analýzy

1. *Volba vstupní databáze*: zadává se typ dat (a) proměnných (sloupců) analyzovaných objektů (řádků), (b) sloupců matice vzdáleností, (c) sloupců korelační matice.
2. *Volba druhu veličin:* zadává se typ užitých veličin v datech, která mohou být (a) intervalová, (b) ordinální, (c) nominální, (d) symetrická binární, (e) asymetrická binární, (f) poměrová.
3. *Název objektů:* zadání pojmenování či jmen jednotlivých objektů, umístěných v řádcích, které se mohou objevit v dendrogramu místo indexů (pořadových čísel) objektů.
4. *Typ shlukovací techniky:* volba metody z možností: jednoduchá průměrová (Average), skupinového průměru, centroidní (Centroid), nejbližšího souseda (Single, Nearest), nejvzdálenějšího souseda (Complete, Furthest), mediánová (Median), Wardova, a flexibilní.
5. *Volí se druh užité vzdálenosti:* vzdálenosti mohou být Eukleidova metrika čili geometrická vzdálenost, Hammingova metrika čili Manhattanská vzdálenost, zobecněná Minkowskiho metrika a Mahalanobisova metrika.
6. *Postup linkování a zařazení do shluků:* tabelární výpočet vzdáleností (nebo podobností) mezi objekty a shluky a postupné vytváření dendrogramu. Postupy jsou (1) metodou hierarchického shlukování, (2) shlukování metodou nejbližších středů, (3) shlukování metodou středů-medoidů, a (4) metodou fuzzy shlukování.
7. *Výpočet skutečných a predikovaných vzdáleností v dendrogramu:* jsou porovnány skutečné vzdálenosti mezi objekty a vypočtené vzdálenosti (predikované) v dendrogramu, jejich rozdíl a konečně i procentuální vyjádření tohoto rozdílu.
8. *Hledání nejlepší techniky tvorby dendrogramu:* dle bodu 4. a 5. lze k sestrojení optimálního dendrogramu kombinovat řadu technik. Rozhodčím kritériem věrohodnosti jsou především kofenetický korelační koeficient *CC*, obě míry těsnosti proložení *delta*, ale také další kritéria: mezishluková suma čtverců WSS_K, procento variace PV_K, silueta *s*, průměrná silueta *SC*, Wilkova statistika λ, rozdělovací koeficienty Dunnův *F(U)* a Kaufmanův *D(U)*.

9. *Vysvětlení nejlepšího dendrogramu podobností objektů:* interpretace optimálního dendrogramu podobnosti jednotlivých objektů je prvním a nejdůležitějším cílem shlukové analýzy.
10. *Vysvětlení nejlepšího dendrogramu podobností proměnných:* interpretace optimál-ního dendrogramu podobnosti jednotlivých proměnných odhalí souvislosti ve struktuře objektů analyzované databáze a je druhým důležitým cílem shlukové analýzy.

(b) Shlukování metodou nejbližších středů (K-Means)

Při vytváření malého počtu shluků z velkého počtu objektů se jeví nejužitečnější shlukovací metodou. Vyžaduje spojité proměnné a především bez odlehlých hodnot. Diskrétní data mohou být rovněž analyzována, ale mohou způsobit problémy.

Princip metody spočívá v rozdělení *n* objektů o *m* proměnných do *k* shluků tak, že mezishluková suma čtverců je přitom minimalizována. Jelikož počet možných uspořádání je enormě veliký, není praktické očekávat vždy nejlepší řešení. Algoritmus nalezne spíše optimum lokální než globální. Je to takové uspořádání shluků, kdy již přemístění objektu z jednoho shluku do druhého nezpůsobí snížení sumy čtverců. Algoritmus pracuje opakovaně, startuje vždy z jiného počátečního uspořádání. Nakonec vybere optimální řešení ze všech možných dosažených uspořádání shluků.

Uživatel zadává počet shluků, jež mají být nalezeny. Pak jsou vytvořeny prostorové shluky nalezením souboru středů shluků tak, že každý objekt je přiřazen do jednoho shluku, načež jsou určeny nové shluky a celý proces se opakuje.

Předpokládejme *n* objektů rozdělených do *k* shluků. Pak *k*-tý shluk obsahuje n_K objektů. Každý objekt je v jednom řádku popsán *m* proměnnými. Chybějící hodnota *i*-té proměnné v *j*-tém řádku u *k*-tého shluku je označena δ_{ijk}. Data x_{ij} jsou předem standardizována a označena z_{ij}.

Počáteční přiblížení ovlivňuje konečné uspořádání shluků. Proto algoritmus pro každý pokus zcela náhodně přiřazuje každý objekt jednomu shluku. Toto uspořádání je pak optimalizováno. Pokus nastartovat proces z rozličných náhodných uspořádání vysoko zvýší pravděpodobnost nalezení globálního optima počtu shluků.

Kritérium věrohodnosti: jde o kritérium těsnosti proložení, které je založeno na srovnání rozličných konfigurací shluků a vychází z *mezishlukové sumy čtverců* WSS_K definované vztahem

$$WSS_K = \frac{nm}{nm - m} \sum_{k=1}^{k} \sum_{i=1}^{m} \sum_{j=1}^{n_k} (1 - \delta_{ijk})(y_{ij} - c_{ik})^2 ,$$

kde c_{ik} je střední hodnota (průměr) *i*-té proměnné v *k*-tém shluku. *Procento variace* (proměnlivosti) je definováno vztahem:

$$PV_K = 100\% \, \frac{WSS_K}{WSS_1} .$$

Chybějící data: lze řídit vypouštění objektů s chybějícími hodnotami proměnných pomocí procenta chybějících hodnot u proměnných. Objekty, které mají více chybějících proměnných než dovolené procento, jsou z další analýzy vypuštěny.

(c) Shlukování metodou středů-medoidů

Medoid, čili střed shluku, je střední objekt, pro který platí, že průměrná vzdálenost k ostatním objektům v tomto shluku je minimální. Je-li požadováno *k* shluků, bude existovat také *k* medoidů. Po nalezení medoidů jsou data klasifikována do shluků vždy okolo nejbližšího medoidu. Medoidy a shluky se vytvářejí na základě *vzdáleností* čili *nepodobností* (dissimilarities).

Proměnné shluku. Druhů proměnných je celá řada: *Intervalové* jsou spojité kladné či záporné, v lineární škále, např. výška, hmotnost, cena, teplota, čas atd. *Ordinální* jsou pořadová čísla stupnice, hodnotící nějakou vlastnost, např. silný nesouhlas (5), nesouhlas (4), neutrální (3), souhlas (2) a silný souhlas (1). *Poměrové* jsou kladné hodnoty, kdy vzdálenost mezi dvěma čísly je stejná, když i jejich poměr je stejný, např. mezi 3 a 30 je stejná jako mezi 30 a 300, chemická koncentrace, intenzita záření, absorbance atd. *Nominální* jsou proměnné, které vyjadřují pouze kvalitu a nikoliv kvantitu, např. PSČ, rasa, barva, název města atd. *Symetrické binární:* mají dvě možnosti, obvykle ano (1), ne (0). *Asymetrické binární:* přítomnost či nepřítomnost zřídka se vyskytující události, kdy nepřítomnost není tak důležitá, např. osoba má jizvu na tváři, a tím je lépe identifikovatelná.

Späthova metoda. Metoda minimalizuje účelovou funkci přemísťováním objektů z jednoho shluku do druhého. Začíná u počátečního uspořádání shluků, algoritmus pak najde lokální minimum inteligentním přesouváním objektů ze shluku do shluku. Jakmile se nepřemístí už žádný objekt, metoda terminuje. Lokální minimum však nemusí být globálním. Aby program překonal toto omezení, zopakuje se několikrát hledání vždy z jiného startovacího uspořádání a nejlepší uspořádání shluků je nakonec bráno za výsledné. Za účelovou funkci se bere celková vzdálenost mezi všemi objekty ve shluku podle vzorce $D = \sum_{k=1}^{K} \sum_{i \in c_K} \sum_{j \in c_K} d_{ij}$, kde K je celkový počet shluků, d_{ij} je vzdálenost mezi i-tým a j-tým objektem a c_K je soubor všech objektů ve shluku k.

Metoda PAM (Partition Around Medoids). Algoritmus se opět pokouší minimali-zovat celkovou vzdálenost D ve dvou krocích:

1. Nalezne se reprezentativní soubor k objektů. První objekt má nejkratší vzdálenost ke všem ostatním objektům, čili představuje *střed*. Pak se *k-1* objektů nachází tak, že hodnota D je co možná nejmenší.
2. Možné alternativy polohy k objektů jsou vybírány iteračním způsobem. Algoritmus vyhledává dosud nezařazené objekty a přemísťuje je tak, aby se hodnota D snižovala. Iterace skončí, jakmile změny nezpůsobí další snížení hodnoty D.

Silueta: poskytuje klíčovou informaci o dobrém a špatném shluku. Hodnota siluety s se vypočte postupem:

1. Objekt i je ve shluku A a má průměrnou vzdálenost a ke všem objektům ve svém shluku. Je-li ve shluku A jediný objekt, je $a = 0$.
2. Sousední shluk B obsahuje objekty, které jsou nejbližší k objektu i ve shluku A a b je průměrná vzdálenost mezi objektem i a všemi objekty ve shluku B.
3. *Silueta s* objektu i se vyčíslí následovně: když shluk A obsahuje pouze jeden objekt, je $s = 0$. Když $a < b$, je $s = 1 - a/b$. Když $a > b$, je $s = b/a - 1$. Když $a = b$, je $s = 0$.

Silueta se vyčíslí pro každý objekt. Hodnota siluety se mění od -1 do +1 a je mírou úspěšné klasifikace do shluků při porovnání vzdáleností uvnitř shluku A se všemi vzdálenostmi objektů nejbližšího souseda B dle pravidla:

1. Je-li s blízko +1, objekt i je dobře klasifikován do shluku A, protože jeho vzdálenosti k ostatním objektům v tomto shluku jsou podstatně kratší než vzdálenosti k objektům nejbližšího souseda B.

2. Je-li s blízko nule, objekt i se nachází kdesi uprostřed mezi shluky A a B, a čistě náhodou byl přiřazen do shluku A.

3. Je-li s blízko -1, objekt i je špatně klasifikován. Vzdálenosti k ostatním objektů ve svém shluku jsou mnohem větší než vzdálenosti k objektům nejbližšího souseda B. Otázkou pak je, proč byl vlastně zařazen do shluku A.

Určení počtu shluků. Přehlednou statistikou je průměrná silueta s, počítaná přes všechny objekty. Tato hodnota sumarizuje jak těsně shlukové uspořádání tato prokládá analyzovaná data. Snadný způsob nalezení správného počtu shluků spočívá v nalezení takového počtu, který maximalizuje průměrnou siluetu. Označme *maximální hodnotu průměrné siluety* všech shluků k symbolem SC a pak budeme rozlišovat následující typy shlukových uspořádání:

SC	Vysvětlení uspořádání do shluků
od 0.71 do 1.00	Silná a dobrá struktura.
od 0.51 do 0.70	Ještě přijatelná struktura.
od 0.26 do 0.50	Slabá struktura, asi umělá. Je třeba najít novou, lepší.
od -1.00 do 0.25	Naprosto nevhodná struktura.

Diagnostika dobrého shlukování. Čárový diagram siluet, uspořádáný podle stoupajícího počtu shluků a hodnoty siluety, dobře prokazuje nejlepší uspořádání shluků. Důležitým kritériem je kladná hodnota siluety s, která by měla být také větší než 0.50. Je-li silueta pro některé shluky menší než 0.50 nebo dokonce záporná, jsou takové shluky nepravděpodobné a měli bychom hledat jiné.

Další analýza struktury objektů. Po úspěšném nalezení počtu shluků a nejlepšího shlukového uspořádání by měla následovat diskriminační analýza, která statisticky testuje, jak dobře byly objekty (řádky) roztříděny do shluků. Testování se provádí pomocí Wilkovy statistiky λ. Vedle diskriminační analýzy jsou sestrojeny různé rozptylové diagramy, ve kterých je počet shluků použit jako důležitá proměnná. Teprve diagramy odhalí a vysvětlí pravý smysl klasifikační analýzy do shluků.

(d) Fuzzy shlukování

Fuzzy shlukování zobecňuje všechny shlukovací metody tím, že umožňuje shlukování jednoho objektu do více než jednoho shluku, zatímco v běžném shlukování je každý objekt členem pouze jednoho shluku. Předpokládejme, že máme K shluků a budeme definovat soubor proměnných m_{i1} , m_{i2} , ..., m_{iK} , které představují pravděpodobnost, že objekt i je klasifikován do k-tého shluku. V běžném shlukovacím algoritmu je jedna z těchto proměnných rovna jedné a zbytek roven nule. To představuje skutečnost, že takový algoritmus klasifikuje každý objekt do jednoho a právě jednoho shluku.

Ve fuzzy shlukování je "účast objektu", čili přítomnost objektu, rozdělena do všech shluků. Proměnná m_{ik} může zde být rovna 1 nebo 0 a suma těchto hodnot musí být rovna 1. Nazveme tento process *fuzzifikací shlukové konfigurace.* Proces má výhodu, že nenutí objekt aby byl zařazen pouze do jediného specifického shluku. Nevýhodou však je, že se zde objevuje mnohem více informací, které musí být vysvětleny. Fuzzy algoritmus minimalizuje účelovou funkci C, která je funkcí neznámých účastí ve shluku a dále funkcí i vzdáleností dle vztahu

$$C = \sum_{k=1}^{K} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} m_{ik}^2 \, m_{jk}^2 \, d_{ij}}{2 \sum_{j=1}^{N} m_{jk}^2} \quad ,$$

kde m_{ik} představuje neznámou účast objektu i v k-tém shluku k a d_{ij} je vzdálenost mezi objekty i a j. Účasti ve shluku jsou předmětem omezení a musí být nezápornými čísly a dále účasti pro jeden objekt musí být v sumě rovny 1. To znamená, že účasti mají stejná omezení, jako by to byly pravděpodobnosti, že individuum patří do jisté skupiny.

Míra věrohodnosti: jedním z nejobtížnějších úkolů ve shlukové analýze je nalezení vhodného *počtu* shluků. Velikost "fuzzifikace" v řešení se dá změřit *Dunnovým rozdělovacím koeficientem*, který představuje míru, jak těsně padne fuzzy řešení na odpovídající *pevné shluky*. Za pevné shluky budeme považovat klasifikaci každého objektu do shluku, který má největší účast. Dunnův rozdělovací koeficient se vyjádří vzorcem

$$F(U) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} m_{ik}^2 \quad .$$

Koeficient leží v intervalu od *1/K* do 1. Hodnoty *F(U) = 1/K* se dosáhne, když všechny účasti jsou rovny *1/K.* Hodnota *F(U)* = 1 platí, když pro každý objekt je účast jednotková a zbytek je roven nule. Dunnův rozdělovací koeficient může být také normován tak, že jeho hodnota se mění od 0 (úplně fuzzy) do 1 (pevný shluk). Normovaná verze má tvar:

$$Fc(U) = \frac{F(U) - (1/K)}{1 - (1/K)} \quad .$$

Další koeficient představuje *Kaufmanův rozdělovací koeficient*

$$D(U) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} (h_{ik} - m_{ik})^2 \quad .$$

Koeficient má hodnotu danou intervalem od *D(U)* = 0 (pevné shluky) do *D(U)* = 1-(1/*K)* (úplně fuzzy). Normovaná verze tohoto koeficientu má tvar

$$Dc(U) = \frac{D(U)}{1 - (1/K)} \quad .$$

Oba normované koeficienty *Fc(U)* a *Dc(U)* poskytují dohromady dobrou indikaci optimálního počtu shluků. celočíselná hodnota K by měla být volena tak, že *Fc(U)* bude nabývat malé a *Dc(U)* velké hodnoty.