Z Wikipedie, otevřené encyklopedie
ikona
Tento článek potřebuje úpravy.
Můžete Wikipedii pomoci tím, že ho vylepšíte. Jak by měly články vypadat, popisují stránky Vzhled a styl, Encyklopedický styl a Odkazy.

Konkrétní problémy: *sjednotit formátování vzorců*

Shluková analýza (též **clusterová analýza**, anglicky cluster analysis) je vícerozměrná statistická metoda, která se používá ke klasifikaci objektů. Slouží k třídění jednotek do skupin (shluků) tak, aby si jednotky náležící do stejné skupiny byly podobnější než objekty z ostatních skupin. Shlukovou analýzu je možné provádět jak na množině objektů, z nichž každý musí být popsán prostřednictvím stejného souboru znaků, které má smysl v dané množině sledovat, tak na množině znaků, které jsou charakterizovány prostřednictvím určitého souboru objektů, nositelů těchto znaků.

Klasifikace shlukovacích metod

Základní členění shlukovacích metod podle cíle je na hierarchické a nehierarchické metody.

1. *Hierarchické shlukování* vytváří systém podmnožin, kde průnikem dvou podmnožin – shluků je buď prázdná množina, nebo jeden z nich. Pokud nastane alespoň jednou druhý případ, je systém hierarchický. Tedy je to jakési větvení, zjemňování klasifikace. K hierarchickému shlukování lze přistupovat ze dvou stran – rozlišujeme přístup *divizní* (vycházíme z celku, jednoho shluku, a ten dělíme) a *aglomerativní* (vycházíme z jednotlivých objektů, shluků o jednom členu, a ty spojujeme). Hierarchické shlukování nabízí více alternativních řešení, výsledek shlukování je pak možné vyjádřit dendrogramem. Tato metoda však není vhodná pro velké datové soubory.
2. *Nehierarchické shlukování* vytváří takový systém, kde jsou shluky disjunktní množiny. Používá se nejčastěji algoritmus *k*-means.

Měření podobnosti objektů

Shluková analýza vychází z podobnosti, resp. vzdálenosti objektů. Její kvantitativní vyjádření je jedním ze základních problémů clusterové analýzy. Existuje mnoho způsobů konstrukce tohoto ukazatele. Výběr vhodné metriky vzdálenosti je zásadní, protože přímo ovlivňuje výsledky analýzy a strukturu nalezených shluků. Mezi nejčastěji používané metody měření vzdáleností patří Eukleidovská vzdálenost, Manhattanská vzdálenost, Minkowského vzdálenost, Kosinová vzdálenost a další. Každá z těchto metrik má své výhody a omezení v závislosti na povaze a dimenzi dat. [1][2]

Výběr správné metriky závisí na charakteru analyzovaných dat. U numerických hodnot se obvykle preferuje eukleidovská nebo Minkowského vzdálenost, zatímco u kategoriálních dat může být efektivnější Jaccardova vzdálenost. Pro textová nebo vysokodimenzionální data je často nejlepší volbou kosinová vzdálenost.[1][3]

Vlastnosti vzdálenosti

Standardními požadavky pro vhodný předpis míry vzdálenosti (metriky) d dvou objektů O_i a O_j jsou:

- nezápornost: $d(O_i, O_j) \geq 0$;
- symetrie: $d(O_i, O_j) = d(O_j, O_i)$;
- shodné objekty by měly mít ukazatel vzdálenosti roven 0: $d(O_i, O_i) = 0$ (zároveň míra podobnosti bude rovna maximální hodnotě, obvykle 1).
- trojúhelníková nerovnost: $d(O_i, O_j) \leq d(O_i, O_h) + d(O_h, O_j)$.

Typy vzdáleností

- metriky – základní je eukleidovská vzdálenost a od ní jsou odvozeny další ukazatele (např. čtverec euklidovské vzdálenosti), další metriky jsou Manhattanská, Čebyševova ad.)
- koeficienty asociace – určeny pro hodnocení podobnosti pro objekty vyjádřené dichotomickými znaky, ukazatele založeny na počtu shod a počtu znaků. Konkrétních ukazatelů je celá řada, některé operují jen s pozitivními shodami (např. koeficient Jaccardův, Russel & Rao, Diceův ad.), některé i s negativními (např. Sokalův, Hamannův). Mohou se

vztahovat k celkovému počtu znaků, k počtu rozdílných případů nebo k různým kombinacím předchozích.
- korelační koeficient – hodí se především pro shlukování proměnných.

Existuje řada dalších způsobů měření vzdálenosti či podobnosti (míry asymetrie, Lambda, kosinus vektorů, chí-kvadrát). Někdy je způsob hodnocení podobnosti/vzdálenosti přímo dán shlukovací metodou. I pokud tomu tak není, je třeba při výběru ukazatele brát v úvahu metodu shlukování a charakter souboru.

Konkrétní příklady

Eukleidovská vzdálenost

Eukleidovská vzdálenost, známá také jako L2 norma, je nejběžnější metrika používaná v shlukové analýze, a to především díky své jednoduchosti a intuitivní interpretaci jako „přímé čáry“ mezi dvěma body v n-rozměrném prostoru. Eukleidovská vzdálenost, pojmenovaná po starořeckém matematikovi Eukleidovi, byla jednou z prvních formálně popsaných metrik. Její aplikace v moderních statistických metodách však začala až s rozvojem analýzy dat v polovině 20. století.[4][5] Matematicky je definována jako:

$$d(x, y) = \sqrt{(\sum(x_i - y_i)^2)}$$

Tato metrika je vhodná pro metrická data, ale citlivá na extrémy (outliery), což může ovlivnit výsledky analýzy.[2][5]

Euklidovská a Manhattanská vzdálenost

Manhattanská vzdálenost

Manhattanská vzdálenost, známá také jako L1 norma, měří vzdálenost jako součet absolutních rozdílů mezi souřadnicemi. Používá se v případech, kdy se předpokládá pohyb pouze ve vertikálním nebo horizontálním směru: [6]

$$d(x, y) = \sum|x_i - y_i|$$

Tato vzdálenost je vhodná pro data s kategoriemi nebo mřížkovou strukturou, například ve městských prostorových analýzách.[2]

Minkowského vzdálenost

Minkowského vzdálenost je zobecněním jak eukleidovské, tak manhattanské vzdálenosti, řízeným parametrem p:

$d(x, y) = (\sum|x_i - y_i|^p)^{1/p}$

Například pro p = 2 získáme eukleidovskou vzdálenost a pro p = 1 manhattanskou vzdálenost. Tento flexibilní přístup je výhodný pro analýzu různorodých dat.[2][4]

Kosinová vzdálenost

Kosinová vzdálenost měří úhel mezi dvěma vektory, což je užitečné při práci s vysokodimenzionálními daty, jako jsou textové dokumenty nebo datové matice. Je definována jako:

$d(x, y) = 1 - (x \cdot y) / (||x|| \, ||y||)$

Kosinová vzdálenost nezohledňuje absolutní hodnoty vektorů, ale zaměřuje se na jejich orientaci.[5][3]

Jaccardova vzdálenost

Používá se především pro binární data, například při analýze přítomnosti či nepřítomnosti určitých vlastností. Definice je následující:

$d(x, y) = 1 - (|x \cap y|) / (|x \cup y|)$

Jaccardova vzdálenost je běžná při analýze genetických dat nebo dat sociálních sítí.[3][6]

Metody hierarchického shlukování

Existují různé přístupy, jak shlukovat objekty na základě jejich vzdálenosti či podobnosti. Mezi základní metody patří:

- *metoda nejbližšího souseda* (*single linkage*, *nearest neighbor*) – vzdálenost shluků je určována vzdáleností dvou nejbližších objektů z různých shluků. Při použití této metody jsou objekty taženy k sobě, výsledkem jsou dlouhé řetězy.

- *metoda nejvzdálenějšího souseda* (*complete linkage*, *furthest neighbor*) – vzdálenost shluků je určována naopak vzdáleností dvou nejvzdálenějších objektů z různých shluků. Funguje dobře především v případě, že objekty tvoří přirozeně oddělené shluky, nehodí se, pokud je tendence k řetězení.
- *centroidní metoda* – vzdálenost shluků je určována vzdáleností jejich center (hypotetická jednotka s průměrnými hodnotami znaků). Může být nevážená nebo vážená. Ta zohledňuje velikosti shluků a je vhodná pokud očekáváme jejich rozdílnost. Užívá se vyjádření vzdálenosti objektů čtvercovou euklidovskou vzdáleností.
- *párová vzdálenost* (pair-group average) – vzdálenost shluků je určována jako průměr vzdáleností všech párů objektů z různých shluků. Opět může být ve vážené i nevážené podobě.
- *Wardova metoda* – vychází z analýzy rozptylu. Slučuje takové shluky, kde je minimální součet čtverců. Obecně lze říci, že je tato metoda velmi účinná, nicméně má tendenci vytvářet poměrně malé shluky. Vzdálenosti objektů se měří čtvercovou euklidovskou vzdáleností.

Příklady využití

Zdravotnictví

Shluková analýza je užitečným nástrojem ve zdravotnictví pro porozumění rozdílům a podobnostem mezi pacienty s Parkinsonovou chorobou (PD), což umožňuje definovat specifické podskupiny. Tyto podskupiny mohou sloužit k přesnější diagnostice, sledování průběhu nemoci a plánování léčby. Výzkumy v této oblasti často využívají klinické škály, jako je například UPDRS, k seskupování pacientů na základě různých symptomů a jejich závažnosti, přičemž výsledky typicky odhalují dvě až pět klastrů. Identifikace těchto podskupin je důležitá pro vývoj personalizované péče, která lépe odpovídá individuálním potřebám pacientů a umožňuje efektivnější léčebné strategie. [7]

Finanční sektor

Banka může pomocí shlukové analýzy rozdělit své klienty do rizikových skupin na základě jejich demografických a socioekonomických charakteristik, včetně kombinace numerických údajů (např. příjem, výše úvěru) a

kategoriálních dat (např. zaměstnání, rodinný stav). Tento přístup umožňuje lépe identifikovat klienty s vysokým rizikem nesplacení úvěru. Například pomocí sjednoceného podobnostního metru mohou být klienti rozděleni do podskupin, kde jedna z nich vykazuje vyšší pravděpodobnost problémů s placením. Na základě těchto výsledků může banka upravit úvěrové limity, přizpůsobit podmínky nebo nabídnout preventivní opatření, jako je finanční poradenství. Tento přístup podporuje efektivnější řízení rizik a minimalizuje potenciální ztráty. [8]

Lesnictví

Lesní správa může pomocí shlukové analýzy efektivně rozdělit lesní stanoviště do skupin na základě jejich ekologických a ekonomických charakteristik, jako jsou náklady na pěstební opatření, typ půdy nebo druhové složení porostu. Tento přístup umožňuje lépe pochopit, jak se jednotlivá stanoviště liší z hlediska nákladové náročnosti a potřebné péče. Například pomocí matice nepodobnosti a její vizualizace lze stanoviště seskupit tak, že některé z nich vykazují vyšší náklady na specifické pěstební zásahy. Na základě těchto výsledků může lesní správa optimalizovat plánování zásahů, přizpůsobit metody péče nebo efektivněji alokovat zdroje. Tento přístup podporuje udržitelnější hospodaření v lesích a snižuje celkové náklady na péči. [9]

Odkazy

Reference

1. ↑ ***a*** ***b*** ŘEZANKOVÁ, Hana; HÚSEK, Dušan a SNÁŠEL, Václav. *Shluková analýza dat*. 2., rozš. vyd. Praha: Professional Publishing, 2009. ISBN 978-80-86946-81-8
2. ↑ ***a*** ***b*** ***c*** ***d*** MELOUN, Milan; MILITKÝ, Jiří a HILL, Martin. *Statistická analýza vícerozměrných dat v příkladech*. Vydání druhé, v Nakladatelství Karolinum první. Praha: Univerzita Karlova v Praze, nakladatelství Karolinum, 2017. ISBN 978-80-246-3618-4
3. ↑ ***a*** ***b*** ***c*** HASTIE, Trevor; TIBSHIRANI, Robert a FRIEDMAN, J. H. *The elements of statistical learning: data mining, inference, and predition*. 2nd ed. Springer series in statistics. New York: Springer Science+Business Media, c2009. ISBN 978-0-387-84857-0

4. ↑ **a b** HEBÁK, Petr; HUSTOPECKÝ, Jiří a MALÁ, Ivana. *Vícerozměrné statistické metody*. [2]. Praha: Informatorium, 2005. ISBN 80-7333-036-9
5. ↑ **a b c** KAUFMAN, Leonard a ROUSSEEUW, Peter J. Finding Groups in Data: An Introduction to Cluster Analysis. 2. vyd. New York: Wiley, 1990. ISBN 0-471-87876-6
6. ↑ **a b** RENCHER, Alvin C. a CHRISTENSEN, William F. *Methods of multivariate analysis*. 3rd ed. Wiley series in probability and statistics. Hoboken: John Wiley, c2012. ISBN 978-0-470-17896-6.
7. ↑ HENDRICKS, Renee M.; KHASAWNEH, Mohammad T. A systematic review of Parkinson's disease cluster analysis research. *Aging and disease*, 2021, 12.7: 1567.
8. ↑ CARUSO, Giulia, et al. Cluster Analysis for mixed data: An application to credit risk evaluation. *Socio-Economic Planning Sciences*, 2021, 73: 100850.
9. ↑ Economic aspects of silvicultural treatments in oak Stands. (2023) *Sylwan*, 167 (12), pp. 828-837.

Literatura

- A. Lukasová, J. Šarmanová: Metody shlukové analýzy. SNTL, Praha 1985.

Externí odkazy

- Obrázky, zvuky či videa k tématu shluková analýza na Wikimedia Commons
- Shluková analýza v České terminologické databázi knihovnictví a informační vědy (TDKIV)
- Detailnější rozbor včetně matematického aparátu [1]
- Kapitola z české online učebnice prostorových analýz [2]
- Kapitola z anglické online učebnice statistiky (anglicky)[3] Archivováno 4. 2. 2007 na Wayback Machine.
- Popis a ukázka aplikace některých shlukovacích algoritmů – fuzzy shluková analýza, shlukování kolem medoidů (PAM) a CLARA – a prostředků k hodnocení klasifikačního modelu (silhouette plot) [4] Archivováno 15. 2. 2010 na Wayback Machine.

Kategorie:

- Data mining
- Matematická statistika

Skryté kategorie:

- Monitoring:Stránky používající kouzelné odkazy ISBN
- Údržba:Články k úpravě
- Monitoring:Články s odkazem na TDKIV
- Monitoring:Články s identifikátorem NKC
- Monitoring:Články s identifikátorem TDKIV
- Monitoring:Články s identifikátorem BNF
- Monitoring:Články s identifikátorem LCCN
- Monitoring:Články s identifikátorem NLI

Hledání

Speciální:Hledání

Hledat

Shluková analýza

40 jazyků

Přidat téma