CLUSTERING

David Bystroň

Fakulta strojního inženýrství, Vysoké učení technické v Brně
Ústav automatizace a informatiky
Technická 2896/2, Brno 616 69, Česká Republika
228676@vutbr.cz

Abstrakt: *Tato semestrální práce se zabýva clusteringem, rozdělením algoritmů na pár základních skupin a stručným popisem některých jednoduchých metod clusteringu jako je K-means, DBSCAN a STING.*

Klíčová slova: *clustering, strojové učení, K-means, DBSCAN, STING, Single Link, Complete Link*

1 Clustering

Clustering je označení pro výpočetní metody, sloužící ke klasifikaci prvků do jednotlivých skupin na základě jejich vlastností nebo určitého stupně podobnosti. Základní myšlenkou je tak zařazení jednotlivých prvků, které jsou si podobné na základě předem definovaných kritérií do stejné skupiny. [3] Tato definice je velmi obecná a metody clusteringu se výrazně liší na základě řešené problematiky. Mezi běžné aplikace může patřit například kolaborativní filtrování, segmentace zákazníků, analýza biologických dat nebo detekci objektů či robotické roje.[1]

2 Rozdělení metod

Algoritmů pro clustering existuje velmi mnoho, lze je však rozdělit do základních skupin podle toho jak algoritmy fungují. Ne všechny algoritmy jde jednoznačně přiřadit jedné skupině a výběr správného algoritmu není také snadný. Neexistuje objektivně nejlepší metoda pro daný problém.[10] Níže budou popsány základní skupiny do kterých můžeme algoritmy dělit.

2.1 Algoritmy částečného shlukování

Algoritmy rozkládají soubor dat na množinu nesouvislých shluků. Klasifikují data do předem daného počtu skupin přičemž každá skupina musí obsahovat alespoň jeden bod a každý bod patří přesně do jedné skupiny.[5] Do této skupiny patří K-means algoritmus a jemu blízké algoritmy jako třeba K-medoids, Fuzzy K-means, Bisecting K-means atd.

2.1.1 K-means clustering

K-mean algoritmus je jeden z nejznámějších učících algoritmů bez učitele. Mějme data set $X = \{x_1, ..., x_n\}$ v d-dimensionálním Euklidovském prostoru. $A = \{a_1, ..., a_c\}$ jsou středy shluků c. A $z = [z_{ik}]_{n \times c}$ kde z_{ik} je buď 0, nebo 1, podle toho zda bod x_{ik} náleží do k-tého shluku, kde $k = 1, ..., c$. Funkce K-means algoritmu je pak

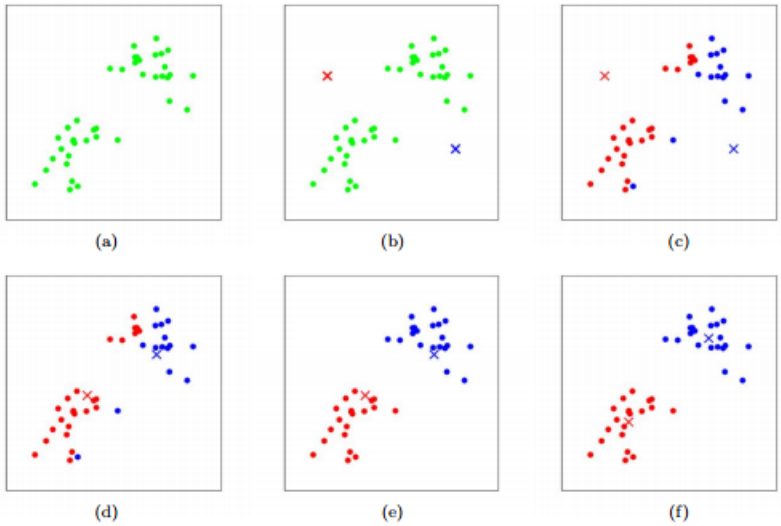
$$J(z, A) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \|x_i - a_k\|^2 . \tag{1}$$

Algoritmus je pak iterován přes nutné podmínky pro minimalizaci funkce $J(z, A)$ pomocí následujících rovnic.

$$a_k = \frac{\sum_{i=1}^{n} z_{ik} x_{ij}}{\sum_{i=1}^{n} z_{ik}} \tag{2}$$

$$z_{ik} = \begin{cases} 1 & \text{if } \|x_i - a_k\|^2 = \min\limits_{1 \le k \le c} \|x_i - a_k\|^2 \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Můžeme zde vidět problém, že je třeba předem určit počet shluků, ten však obvykle není znám.[9]



Obrázek 1: Postup center shluků a přiřazování bodů K-means algoritmu[7]

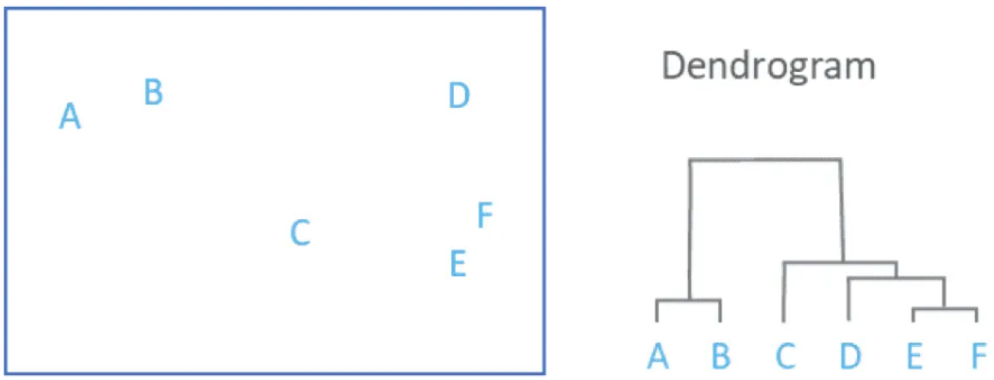
2.2 Hiearchický clustering

Hiearchické algoritmy byly vyvinuty za účelem determinističtějšího a flexibilnějšího mechanismu pro clustering. Můžeme je rozdělit na aglomerativní a dělící metody.

Aglomerativní začínají tím, že vezmou jednotlivé shluky, kde každý obsahuje jen jeden objekt, na nejnižší úrovni. Pak pokračují slučováním dvou shluků a vytvoří hiearchii zdola nahoru.

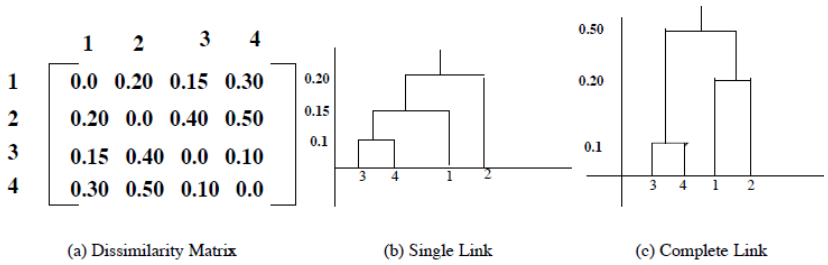
Dělící metody naopak začínají se všemi datovými objekty v jednom obrovském shluku a postupně shluky rozděluje na dvě skupiny, čímž vytvoří hiearchii od shora dolů.[1]

Hlavním výstupen algoritmu je pak dendogram, který znázorňuje hierarchii a vztahy mezi jednotlivými shluky.



Obrázek 2: Dendogram hiearchickégo algoritmu[2]

Mezi nejznámější algoritmy patří Single Link, kde podobnost dvou shluků je podobnost mezi nejpodobnějšími členy. A Complete link, kde se meří podobnost dvou shluků jako podobnost dvou nejodlišnějších členů. Další algoritmy jsou pak například CHAMELEON nebo COBWEB.[1]



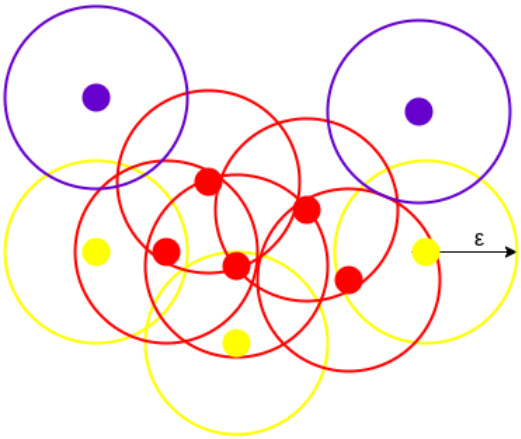
Obrázek 3: Aglomerativní algoritmy[1]

2.3 Clustering na základě hustoty

Algoritmy založené na hustotě hledají shluky pomocí množství datových bodů v určitém prostoru. Hlavní výhodou těchto metod je to, že jelikož zkoumají prostor na vysoké úrovni granuality, lze je použít k rekonstrukci celého tvaru rozložení dat. Problémem algoritmů bývá rostoucí obtížnost výpočtů ve více dimenzích. Nejpoužívanější algoritmy jsou DBSCAN, DENCLUE a OPTICS.[1]

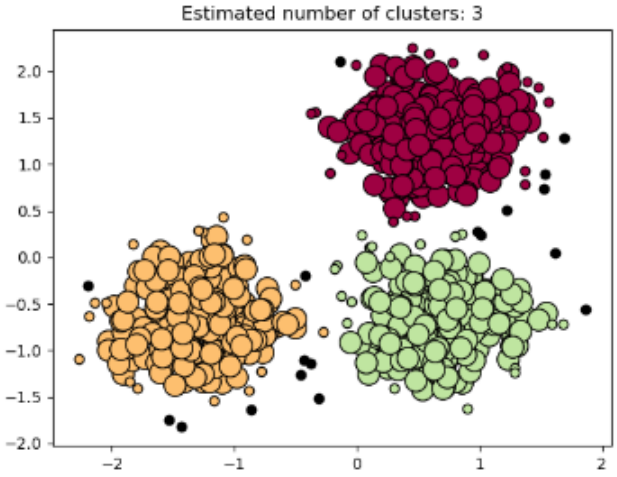
2.3.1 DBSCAN

Pro jednoduchou představu bude DBSCAN popsán na 2D datové množině. V této situaci jsou vstupní parametry algoritmu ε, určující poloměr okolí bodu, a minimální počet bodů, označené jako $minPoints$. Algoritmus funguje tak, že vybere libovolný bod a pokud se v jeho okolí daném radiusem ε nachází alespoň minimální počet bodů, včetně vybraného, je bod označen jako jádro. V případě, že se v okolí nachází menší počet bodů, je bod označen jako hranice. Když se kolem bodu nenachází žádný další bod, je označen jako šum. Na Obr. 4 jsou jádra vyznačeny červeně, hranice žlutě a šum je označen modrou barvou, $minPoints = 3$.[6]



Obrázek 4: Popis DBSCAN[8]

Následně se vybere libovolné jádro a přiřadí se do prvního shluku, všechny ostatní jádra v jeho okolí se rovněž přiřadí do stejného shluku. Tak se rozdělí všechny body označené jako jádro. Následně se přidají ke shlukům body označené jako hranice. Pokud je hraniční bod na pomezí dvou shluků, bude spadat do toho, který se vyhodnocoval jako první. Výsledek pak může vypadat jako na Obr. 5.[6]



Obrázek 5: Výsledek DBSCAN[6]

2.4 Clustering na základě mřížky

Algoritmy na základě mřížky jsou účinné při práci s více rozmětnými daty. Algoritmy rozdělí datový prostor na konečný počet buněk, tím vytvoří mřížkovou strukturu.Poté vytvoříte shluky z buněk v mřížce. Díky tomu, že počet buněk mřížky je mnohem menší než počet datových objektů, velikou výhodou těchto metod je jejich časová náročnost. Algoritmy se zpravidla řídí těmito pěti kroky[1]:
1. Vytvoření struktury mřížky, tj. rozdělení datového prostoru na konečný počet buněk.
2. Výpočet hustoty buněk pro každou buňku.
3. Seřazení buněk podle jejich hustoty.
4. Identifikace center shluků.
5. Procházení sousedních buněk.

Mezi klasické metody patří GRIDCLUS, BANG nebo dále popsaný STING. K vícedimenzionálním se řadí třeba CLIQUE.

2.4.1 STING

V metodě STING je prostor rozdělen na obdelníkové buňky a několik úrovní buněk s různými úrovněmi rozlišení. Buňky vysoké úrovně jsou rozděleny na buňky nižší úrovně. Statistické informace o atributech v každé buňce, jako je průměr, maximální a minimální hodnota, jsou předem vypočteny a uloženy jako statistické parametry. Tyto statistické parametry jsou užitečné pro zpracování dotazů a další úlohy analýzy dat.[4][1]

Algoritmus pro tuto metodu je pak následující[1]:
1. Určete úroveň, na které začnete.
2. Pro každou buňku této úrovně vypočítáme interval spolehlivosti (nebo odhadovaný rozsah) pravděpodobnosti, že tato buňka je relevantní pro dotaz.
3. Z výše vypočteného intervalu označíme buňku jako relevantní nebo nerelevantní.
4. Pokud je tato úroveň listovou úrovní, přejděte ke kroku 6; v opačném případě přejděte ke kroku 5.
5. Přejděte o jednu úroveň níže ve struktuře hierarchie. U buněk, které tvoří příslušné buňky vyšší úrovně, přejděte ke kroku 2.
6. Pokud je splněna specifikace dotazu, přejděte ke kroku 8; v opačném případě přejděte ke kroku 7.
7. Získejte ta data, která spadají do příslušných buněk, a proveďte další zpracování. Vraťte výsledek který splňuje požadavek dotazu. Přejděte na krok 9.
8. Najděte oblasti relevantních buněk. Vraťte ty oblasti, které splňují požadavek dotazu. Přejděte na krok 9.
9. Ukončete algoritmus

3 Závěr

V této práci byl objasněn pojem clustering. Metody clusteringu mají velmi široké uplatnění napříč různými odvětvími, a proto jich existuje poměrně velké množství. Z tohoto důvodu je zde i základní rozdělení metod

podle to jak fungují, nejsou však uvedeny všechny možnosti jak lze algoritmy dělit, z důvodu obsáhlosti této problematiky. V práci jsou také pospány i konkrétní, nejznámější algoritmy K-means, Single a Complete Link, DBSCAN, STING.

Z hlediska robotiky lze clustering použít při řešení robotích rojů, plánování cest nebo detekci objektů pomocí dat ze senzorů jako například lidar.

Reference

[1] Aggarwal, C. C., and Reddy, C. K. *Data Clustering: Algorithms and Applications*, 1 ed., vol. 31 of *Chapman & Hall/CRC data mining and knowledge discovery series*. CRC Press, Milton, 2014.

[2] Bock, T. What is hierarchical clustering? *DisplayR*.

[3] Cruz, N. B., Nedjah, N., and de Macedo Mourelle, L. Robust distributed spatial clustering for swarm robotic based systems. *Applied Soft Computing 57* (2017), 727–737.

[4] GeeksforGeeks. Sting – statistical information grid in data mining, 2022.

[5] Jin, X., and Han, J. *Partitional Clustering*. Springer US, Boston, MA, 2010, pp. 766–766.

[6] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[7] Piech, C. K means. *Standford CS221* (2013).

[8] Sharma, A. How to master the popular dbscan clustering algorithm for machine learning. *Analytics Vidhya* (2024).

[9] Sinaga, K. P., and Yang, M.-S. Unsupervised k-means clustering algorithm. *IEEE Access 8* (2020), 80716–80727.

[10] Wikipedia contributors. Cluster analysis — Wikipedia, the free encyclopedia, 2024. [Online; accessed 24-February-2024].