Konsenzuálne zhlukovanie

Lukáš Šníder

1 Úvod

S exponenciálnym nárastom objemu dát v nedávnej dobe prišli aj vyššie nároky na ich spracovanie a využitie ich skrytého potenciálu. Rôzne techniky orientujúce sa na spracovanie veľkých objemov dát sa tešia čoraz väčšej popularite, patrí medzi ne napríklad zhluková analýza. Vo všeobecnosti môžme tento prístup zjednodušiť na rozdeľovanie dát do skupín, ktoré sú určitým spôsobom zmysluplné alebo poskytujú podklad pre ďalšie spracovanie.

V tejto práci sa bližšie pozrieme na základné techniky spracovania dát v rámci zhlukovej analýzy, ich výhody a nevýhody s dôrazom na možnosti ich uplatnenia.

2 Zhluková analýza

Zhluková analýza je jednou z techník dátovej analýzy, ktorej cieľom je nájsť v analyzovaných dátach ich podmnožiny, v ktorých je možné sledovať známky podobnosti medzi jednotlivými prvkami každej takejto podmnožiny a zároveň medzi prvkami rôznych podmnožín sú viditeľné rozdielnosti. Takto identifikované podmnožiny sa nazývajú zhluky (z angl. clusters). [Russell and Norvig, 2003]

Výstupy zhlukovej analýzy - zhluky a priraďovanie dátových objektov k nim môžu nabádať k zaradeniu tohto prístupu medzi ostatné klasifikačné techniky, pri ktorých máme k dispozícií dopredu definované triedy, do ktorých už iba zaraďujeme skúmané dátové množiny. Nakoľko ale proces zhlukovania derivuje štruktúru výsledných zhlukov čisto z prístupných dát, tento prístup sa zaraďuje medzi tzv. "učenia bez učiteľa" (z angl. unsupervised learning), pri ktorých nemáme k dispozícií dopredu definované označené dáta alebo triedy dát, do ktorých by bolo možné skúmané dáta zaradiť. [Tan et al., 2005]

Absencia potreby dodatočných informácií o skúmaných dátach pred samotnou analýzou poskytuje priestor pre jej široké uplatnenie, čo aj dokumentuje prax. Zhlukovanie sa podarilo úspešne aplikovať v odvetviach ako rozpoznávanie obrazu, prehliadanie webu, business intelligence, dolovanie dát, biológia, analýza klimatických údajov, psychológia a medicína, atď. [Han, 2005]

2.1 Typy zhlukov

Jasne Oddelené (z ang. well-separated) zhluky sú také, v ktorých sú dátové objekty rozdelené takým spôsobom, že v každom zhluku sú si v rámci ich zhluku

navzájom podobnejšie, resp. sú k sebe bližšie ako ku každému inému objektu, ktorý sa nachádza mimo ich zhluku. V podstate môžu nadobúdať akékoľvek tvary. **Prototypové** (z ang. prototype-based) tiež nazývané aj centralizované (center-based) sú špecifické tým, že je možné za jednotlivé zhluky identifikovať ich reprezentantov (prototypy), ktorí predstavujú reprezentatívny bod daného zhluku, v závislosti na tom akú zhlukovaciu metódu sme pri ich tvorbe zvolili môže ísť o centroidy, medoidy, atď. Tieto typy zhlukov nadobúdajú kruhové tvary. **Grafovo-založené** (z ang. graph-based) sú špecifické tým, že sa na ne môžme pozerať ako na prepojenú štruktúru, kde môžu byť vnútorné spojenia definované ako špecifikovaná vzdialenosť medzi dvoma dátovými objektmi vďaka čomu môžu nadobúdať neštandardné tvary. **Intenzitou definované** (z ang. density-based) sa vymedzujú ako zhluky, ktoré sú definované oblasťami s vysokou intenzitou výskytu dátových objektov a sú ohraničené s oblasťami s pozorovateľne nižšou intenzitou výskytu dátových bodov. Hľadanie týchto typov zhlukov sa hodí v situáciách, keď pracujeme s nekvalitnými dátami s množstvom šumu, zhluky nie sú jasne definované a sčasti navzájom prepletené. **Koncepčné** (z ang. conceptual) zhluky, tiež nazývané zhluky so spoločnou vlastnosťou obsahujú dátové objekty s konkrétnou črtou, ktorá je spoločná pre všetky prvky konkrétneho zhluku. Keďže takto definovaná podmienka na začlenenie dátového objektu do zhluku je relatívne slabá, jeden objekt sa môže ocitnúť ako člen viacerých hlukov. Tieto typy zhlukov častokrát odhalia skryté vzory v dátach. [Tan et al., 2005]

2.2 Metódy zhlukovej analýzy

Rozdeľovacie (partitioning) metódy možno vymedziť ako prístupy, ktoré sa snažia jednotlivé dátové objekty zadeliť do disjunktných pomnožín na základne zvolenej miery vzdialenosti. Typicky prebieha tento proces vo viacerých iteráciách, v prvej iterácií sa náhodne vyberú iniciálne rozdelenia dát na základe vzdialenosti k zvolenému stredu rozdelenia (zhluku) a v každom ďalšom behu sa pokúšame tieto rozdelenia vylepšiť, čo v tomto prípade znamená hľadanie alternatívneho priradenia niektorých dátových objektov za účelom definovania rozdelení (zhlukov) s vyššou mierou podobnosti, alebo inak povedané s dátovými objektmi s menšími vzájomnými vzdialenosťami k centrálnym bodom jednotlivých zhlukov. [Han, 2005]

Hierarchické (hierarchical) metódy možno rozdeliť na dve skupiny. Prvé nazývané "top-down" sa snažia vstupné dáta postupným procesom rozkladania zadeliť do špecifických zhlukov, ak máme definovú nejakú ukončovaciu podmienku, alebo sa postupných rozdeľovaním množiny všetkých vstupných dát dostať až k jednotlivým objektom. Druhá skupina "bottom-up" postupuje od opačného konca a teda začína s jednotlivými dátovými objektmi ako zhlukmi samými o sebe a postuným zlučovaním objektov vytvára ďalšie zhluky až dovtedy, kým nedosiahneme ukončovaciu podmienku, prípadne keď tá nie je definovaná nie sú všetky objekty v jednom zhluku. [Zaki and Wagner Meira, 2014]

Mriežkové (grid-based) metódy spočívajú v rozdelení dátového priestoru do menších častí na základe mriežky, podľa ktorej následne prebiehajú všetky

prepočty priraďujúce objekty k zhlukom. [Han, 2005]

Hustotou-orientované (density-based) metódy sa vyznačujú priraďovaním dátových objektov ku konkrétnemu zhluku na základe početnosti ich výskytu v blízkosti stredu tohto zhluku [Han, 2005]

3 Miery podobnosti a nepodobnosti

Miery podobnosti a nepodobnosti slúžia ako ich názov naznačuje na identifikáciu miery toho ako podobné resp. rozdielne sú skúmané objekty medzi sebou. Miera je v tomto prípade typicky reprezentovaná číselnou hodnotou z intervalu (0,1). Priradenie týchto hodnôt nám následne umožní jednotlivé objekty zaradiť do konkrétnych zhlukov, na základe ich skóre z tohto vzájomného porovnania, tým spôsobom, že v rámci jedného zhluku sa nachádzajú objekty, ktoré sú medzi sebou podobné a zároveň v porovnaní s ostatnými objektmi rozdielne. [Han, 2005]

3.1 Meranie miery podobnosti

Pre správne priradenie vzájomných podobností medzi jednotlivými dátovými objektmi a ich správne priradenie do zhlukov s vysokou mierou podobnosti v rámci objektov je potrebné si zvoliť spôsob, akým sa tieto miery budú vyhodnocovať. Pri tomto výbere musíme zohľadniť druh atribútov, ktoré sa snažíme analyzovať, pre numerické atribúty sa štandardne používa Euklidovská alebo Manhattanská vzdialenosť, zatiaľ čo pri kategorických atribútoch siahneme zrejme po kosínusovskej vzdialenosti

3.2 Typy vzdialeností

Najpoužívanejším typom vzdialenosti, ktorý sa pri dátovej analýze používa je **Euklidovská vzdialenosť**. Keď sa na ňu bližšie pozrieme zistíme, že v podstate reprezentuje našu intuitívnu predstavu o vzdialenosti, je to teda priama čiara medzi dvomi bodmi v štandardom priestore na aký sme zvyknutí z každodenného života a vieme ju získať z nasledovného vzorca:

$$d(i,j) = \sqrt{(x_{i1}^2 - x_{j1}^2)^2 + (x_{i2}^2 - x_{j2}^2)^2 + ... + (x_{ip}^2 - x_{jp}^2)^2} \quad (1)$$

Ďalším používaným typom vzdialenosti je **Manhattanská vzdialenosť (vzdialenosť mestských blokov)**:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}| \quad (2)$$

Ako možno vidieť zo vzorca ide o súčet vzdialeností medzi dvoma bodmi všetkých rozmerov priestoru, v ktorom sa dané body nachádzajú.

Kosínusovská vzdialenosť je častou voľbou pri porovnávaní podobnosti dvoch dokumentov. V predspracovaní vytvoríme vektor s frekvenciou výskytu jednotlivých výrazov v dokumente, teda každý dokument bude reprezentovaný

vektorom, ktorého hodnoty budú reprezentovať slová zjednotenia slov všetkých porovnávaných dokumentov

$$sim(x, y) = \frac{\mathbf{x}.\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} \tag{3}$$

Pred

$$d(i, j) = \tag{4}$$

[Han, 2005]

3.3 K-means zhlukovanie

K-means popis... [MacQueen et al., 1967]

3.4 Alternatívy ku K-means

K-medians nejaky popis...

References

Jiawei Han. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005. ISBN 1558609016.

James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2 edition, 2003. ISBN 0137903952.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005. ISBN 0321321367.

Mohammed J. Zaki and Jr. Wagner Meira. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, May 2014. ISBN 9780521766333.