UDK: 004.62:519.237.8 Ekonomski izazovi/ Economic Challenges, Year 13, no. 25, pg. 20-31
004.62:303.722.4 doi: 10.5937/EkoIzazov2425020A / Date of receiving/ Dan primanja: 19.04.2023.
Originalni naučni rad / Original science paper - Date of acceptance/Dan prihvatanja: 02.11.2023.

BIG DATA KLASTER ANALIZA

Melisa Azizović
Doktorand na Departmanu za računarske nauke Univerziteta u Novom
melisa.azizovic@gmail.com, ORCID: 0000-0002-6969-5133

Apstrakt
U ovom radu smo se fokusirali na analizu grupisanja podataka kao najčešće korišćene tehnike za grupisanje različitih objekata. Grupisanjem podataka, možemo izdvojiti grupe sličnih objekata iz različitih kolekcija. Prvo smo definisali Big data i klastering kako bismo pratili dalji sadržaj rada. Predstavili smo najpopularnije tehnike grupisanja podataka, uključujući particionisanje, hijerarhijsko grupisanje, grupisanje na osnovu gustine i grupisanje zasnovano na mreži podataka. Big Data opisuje velike količine podataka. Visoka preciznost velikih podataka može doprineti samopouzdanju u donošenju odluka, a bolje procene mogu pomoći u povećanju efikasnosti, smanjenju troškova i rizika. Za obradu podataka koriste se različite metode i pristupi, uključujući grupisanje, klasifikaciju, regresiju, veštačku inteligenciju, neuronske mreže, pravila asocijacije, stabla odlučivanja, genetske algoritme i metod najbližeg suseda. Klaster predstavlja skup objekata iz iste klase, što znači da se slični objekti grupišu zajedno, a različiti objekti grupišu odvojeno. Opisali smo K-means algoritam, hijerarhijsko grupisanje, grupisanje zasnovano na gustini - DBSCAN algoritam i STING algoritam za mrežu podataka.

Ključne riječi: Big Data, grupisanje, klastering, K-means algoritam, hijerarhijsko grupisanje, DBSCAN, STING.

BIG DATA CLUSTER ANALYSIS

Abstract
In this paper, we focused on cluster analysis as the most commonly used technique for grouping different objects. By clustering data, we can extract groups of similar objects from different collections. First, we defined Big Data and clustering to follow the rest of the paper. We presented the most popular techniques for clustering data, including partitioning, hierarchical clustering, density-based clustering, and network-based clustering. Big Data describes large amounts of data. High precision of big data can contribute to decision-making confidence, and better estimates can help increase efficiency, reduce costs, and risks. Various methods and approaches are used for data processing, including clustering, classification, regression, artificial intelligence, neural networks, association rules, decision trees, genetic algorithms, and the nearest neighbor method. A cluster represents a set of objects from the same class, which means that similar objects are grouped together, and different objects are grouped separately. We described the K-means algorithm, hierarchical clustering, density-based clustering - DBSCAN algorithm, and the STING algorithm for network-based clustering.

Key word: Big Data, clustering, cluster analysis, K-means algorithm, hierarchical clustering, DBSCAN, STING.
JEL codes: C80

UVOD

U poslednjoj deceniji, Big Data je postao jedan od najvećih izazova u informacionoj tehnologiji. Ubrzani razvoj tehnologije i sve veća dostupnost podataka doveli su do povećanja količine podataka koje organizacije generišu i sakupljaju iz različitih izvora. Međutim, analiza velikog broja podataka predstavlja veliki izazov, a jedan od ključnih problema je pronalaženje relevantnih informacija i razumevanje skrivenih veza među podacima. U tom smislu, analiza klastera je postala jedan od najefikasnijih pristupa u Big Data analizi, jer omogućava grupisanje sličnih objekata u određene kategorije, što olakšava njihovo razumevanje i klasifikaciju. Grupisanje podataka podrazumeva proces klasifikovanja sličnih objekata u istu grupu, dok se različiti objekti stavljaju u različite grupe. Veliki podaci predstavljaju izazov za sistem za upravljanje relacionim bazama podataka, desktop statistike i alate za vizualizaciju (Hájek et al.,1996).
Cilj ovog rada je da prikaže značaj analize klastera u Big Data analizi, sa naglaskom na primenu u različitim oblastima. Takođe cilj rada je da naglasi značaj klastiranja u analizi velikih grupa podataka, kao i da predstavi različite tehnike klastiranja. Grupisanje particionih grupa podataka u klastere je jedna od tehnika, a najpopularniji algoritam za tu metodu je K-means algoritam. U hijerarhijskom grupisanju se koristi dendrogram da bi se ilustrovala hijerarhija klastera, a postoje dve tehnike za odlučivanje da li će klasteri biti spojeni ili podeljeni - grupisanje odozgo prema dole i odozdo prema gore. Gustinu, oblik i količinu podataka koriste mreže podataka koje koriste grupisanje zasnovano na gustini, a najpoznatiji algoritm za ovu metodu je DBSCAN dok se za podatke mreže koristi STING algoritam. U drugom i trećem delu rada se objašnjava pojam Big Data i analiza klastera. Četvrti deo rada opisuje najzastupljenije tehnike klastiranja, dok se u petom delu iznose zaključna razmatranja. U današnjem dobu velikih grupa podataka koje se prikupljaju svakodnevno, algoritmi grupisanja su od velikog značaja za njihovu analizu.

POJAM BIG DATA

Big Data predstavlja velike, složene i heterogene skupove podataka koji prevazilaze kapacitete i mogućnosti tradicionalnih baza podataka za njihovo efikasno obrađivanje i analiziranje (Dumbill, 2012). Ovi podaci dolaze iz različitih izvora, uključujući senzore, web stranice, društvene mreže, mobilne uređaje, transakcije i mnoge druge izvore. Obrada velikih količina podataka zahteva napredne tehnologije i alate poput Hadoop-a, Spark-a, NoSQL baza podataka, tehnika mašinskog učenja i analitike podataka. Big Data se koristi u različitim oblastima, uključujući poslovnu analitiku, zdravstvo, sigurnost, marketing, finansije i druge industrije (Brown et al., 2013).
Big Data karakterišu tri glavne karakteristike, poznate kao "3V": velika količina podataka, visoka raznovrsnost podataka i brzina prikupljanja podataka.Osim ovih karakteristika, postoje i druge karakteristike, kao što su (Higdon et al., 2013):

1. Vrednost: Big Data sadrži značajne informacije na osnovu kojih se mogu stvarati vrednosti u različitim oblastima i sektorima.

2. Verodostojnost: Big Data može sadržati nepouzdane i nevaljane podatke, zbog čega je važno osigurati da se podaci proveravaju i validiraju pre upotrebe.
3. Varijabilnost: Big Data može sadržati podatke različitih formata, struktura i izvora, zbog čega je potrebno koristiti različite tehnologije i alate za obradu i analizu podataka.
4. Zapremina: Big Data se pre svega odnosi na ogromne količine podataka koje su previše velike za tradicionalne baze podataka.
5. Brzina: Big Data se često prikuplja u realnom vremenu, a brzina prikupljanja podataka može biti vrlo visoka, što zahteva brzu obradu i analizu podataka.
6. Varijabilnost brzine: Brzina prikupljanja podataka može se menjati, a podaci se mogu prikupljati brže ili sporije u različitim vremenskim periodima.
7. Varijabilnost kvaliteta: Big Data može sadržati podatke različitog kvaliteta, a kvalitet podataka može se menjati u zavisnosti od izvora podataka.
8. Vizualizacija: Big Data se često vizualizuje kako bi se lakše razumeli i interpretirali podaci, što zahteva korišćenje posebnih alata za vizualizaciju podataka.

Sve ove karakteristike zahtevaju napredne tehnologije i alate za obradu i analizu podataka kako bi se obezbedilo da Big Data bude efikasno i uspešno iskorišćena u poslovnom okruženju.

Big Data se sastoji od nekoliko ključnih komponenti koje su potrebne za njegovu obradu i analizu. Neke od tih komponenti uključuju (Marz & Warren, 2014):

a) Podaci: Ovo je osnovna komponenta Big Data, koja predstavlja velike, heterogene i različite skupove podataka koje je potrebno obraditi i analizirati.
b) Skladište podataka: Big Data skladište podataka predstavlja prostor gde se podaci čuvaju i obrađuju. Ovo skladište može se sastojati od različitih tehnologija i alata za upravljanje podacima, kao što su Hadoop, NoSQL baze podataka, kao i tradicionalne baze podataka.
c) Analitički alati: Analitički alati su softverski alati i tehnologije koji se najčešće koriste za obradu i analizu Big Data. Ovi alati uključuju različite modele analize podataka, kao što su mašinsko učenje, statistička analiza, vizualizacija podataka i druge tehnike.
d) Infrastruktura: Infrastruktura Big Data uključuje hardverske i softverske komponente koje su potrebne za obradu i analizu Big Data. Ove komponente uključuju velike količine računarske snage, kao i različite mrežne tehnologije koje omogućavaju brz protok podataka.
e) Sigurnost: Sigurnost je takođe važna komponenta Big Data. Budući da su ovi podaci često osetljivi i privatni, potrebno je osigurati da su podaci zaštićeni od zloupotrebe, krađe i drugih sigurnosnih pretnji.

Sve ove komponente su neophodne za efikasnu obradu i analizu Big Data, a njihova integracija omogućava da se podaci uspešno koriste za različite poslovne svrhe.

ANALIZA KLASTERA

Grupisanje ili analiza klastera je tehnika za lociranje skupova povezanih podataka u skupu podataka. Jedna od najčešće korišćenih metoda grupisanja u nauci o podacima je ova. Entiteti svake grupe su relativno sličniji jedni drugima nego entitetima u drugim grupama. Proces grupisanja uključuje grupisanje populacije ili tačaka

podataka u više grupa tako da su tačke podataka unutar svake grupe sličnije jedna drugoj od tačaka podataka u drugim grupama. Jednostavno rečeno, cilj je sortirati u klastere sve grupe objekata koji dele slične karakteristike (Aggarwal & Yu, 1998).
To čemo objasniti na primeru. Ukoliko vlasnik kompanije za iznajmljivanje želi da zna šta njegovi klijenti žele kako bi mogao da razvija svoju kompaniju. Nemoguće je da analizira specifičnosti svakog kupca i da za svakog od njih smisli poseban poslovni plan. Ali, može koristiti drugačiji pristup za svaku npr. od pet grupa kupaca tako što će sve svoje kupce grupisati u, recimo, pet grupa na osnovu njihovih obrazaca kupovine. Pojavu nazivamo klasterizacijom.
Klaster je skup predmeta iz iste klase. Drugim rečima, predmeti koji su slični skupljaju se zajedno u jedan klaster, dok su oni koji su različiti grupisani su zajedno u drugi. Pretvaranje gomile apstraktnih stvari u klase povezanih objekata je proces grupisanja.
Odlike klastera su (Olfa & Chiheb-Eddine, 2019):
1. Može se rukovati kolekcijom objekata podataka kao jednom grupom.
2. Kada se radi klaster analiza, skup podataka se prvo deli u grupe na osnovu sličnosti tj. na osnovu toga koliko su podaci slični, a grupisanjima se zatim dodeljuju oznake.
3. Osnovna prednost grupisanja u odnosu na kategorizaciju je njegova sposobnost da se prilagodi promenama i identifikuje ključne karakteristike koje pomažu u razdvajanju različitih grupa.

Tipovi grupisanja

Grupisanje se može široko klasifikovati u dve podgrupe (Han et al., 2012):

- Čvrsto grupisanje: Sa čvrstim klasterisanjem, svaka tačka podataka u potpunosti ili delimično pripada klasteru. Na primer, u gornjem primeru, svaki klijent je dodeljen jednoj od 5 grupa.
- Meko grupisanje: Sa mekim klasterisanjem, određuje se verovatnoća da se svaka tačka podataka nalazi u svakom klasteru umesto da se svaka tačka podataka smešta u poseban klaster. Na primer, na osnovu gore pomenute situacije, svakom kupcu je data verovatnoća da bude u jednom od pet klastera maloprodaje.

Primena klasteringa

Kao što smo već naglasili, grupisanje ima širok opseg primene. Mnoge aplikacije, poput istraživanja tržišta, prepoznavanja obrazaca, analize podataka i obrade slike, koriste analizu klastera. Klasteri imaju brojne primene. Na primer, grupisanje može pomoći trgovcima da identifikuju jedinstvene grupe klijenata na osnovu njihovih kupovnih navika. Takođe se može koristiti u biologiji za stvaranje taksonomija biljaka i životinja, grupisanje gena sa sličnim funkcijama i razumevanje strukturnih karakteristika populacija. Grupisanje takođe pomaže u identifikaciji uporedivih regiona korišćenja zemljišta na osnovu slika sa satelita. Takođe je korisno za klasifikaciju naselja u gradu na osnovu lokacije, vrednosti i tipa stanovanja.
Grupisanje je takođe korisno (Bandyopadhyay & Saha, 2013) u cilju kategorizacije dokumenata na internetu radi pronalaženja informacija. Aplikacije za otkrivanje nepravilnosti, kao što je identifikacija prevare sa kreditnim karticama, takođe koriste grupisanje. Takođe se koristi u analizi društvenih mreža, grupisanju rezultata pretrage, medicinskim snimanjima i drugim oblastima.
Modeli algoritama za klastiranje

Kako je grupisanje subjektivna aktivnost, postoji više mogučnosti da se ovaj cilj postigne. Svaki pristup ima svoj skup smernica za određivanje koliko su dve tačke podataka uporedive. Zaista postoji više od 100 algoritama za grupisanje. Ipak, samo nekoliko algoritama se široko koristi, koje u daljnem tekstu analiziramo (Tan et al., 2005):

- Modeli povezivanja: Kao što njihovo ime implicira, ovi modeli su zasnovani na ideji da su tačke podataka koje su najbliže jedna drugoj u prostoru podataka uporedivije jedna sa drugom od onih koje su udaljenije. Metod hijerarhijskog grupisanja i njegove varijacije su primeri ovih modela.
- Centroidni modeli (particijsko grupisanje) su iterativne metode grupisanja u kojima je stepen sličnosti između dve tačke podataka određen njihovom blizinom centroidu klastera. Popularna metoda koja odgovara ovom opisu je grupisanje K-sredina. Od ključne je važnosti imati prethodno znanje o skupu podataka u ovim modelima jer se broj klastera potrebnih pri zaključku mora prethodno navesti. Da bi se locirao lokalni optimum, ovi modeli se ponavljaju.
- Modeli gustine: Ovi modeli skeniraju prostor podataka za regione sa različitim gustinama tačaka podataka. Odvaja brojne različite zone gustine i grupiše stavke podataka koje spadaju u njih. DBSCAN i OPTICS su uzorni modeli gustine.
- Modeli zasnovani na mreži: U ovom slučaju, objekti zajedno formiraju mrežu. Prostor objekta je kvantovan u konačan broj ćelija koje formiraju mrežnu strukturu. Glavna prednost ove metode je brzo vreme obrade. Takođe jedna od prednosti je što zavisi samo od broja ćelija u svakoj dimenziji u kvantizovanom prostoru. STING algoritam je metoda grupisanja zasnovana na mreži podataka.

ALGORITMI KLASTIRANJA

Analitika velikih podataka klasifikuje ili grupiše uporedive objekte na osnovu njihovih zajedničkih karakteristika koristeći tehniku grupisanja. Ovaj pristup se koristi za istraživanje i pronalaženje skrivenih obrazaca i struktura u ogromnim skupovima podataka. K-srednje vrednosti, hijerarhijsko grupisanje, DBSCAN i STING su neke od najčešće korišćenih tehnika grupisanja, koje će biti detaljno analizirane u daljnem tekstu.

K-means algoritam

Najpopularniji algoritam za particiono klasterovanje je K-means (K-sredina). K-means algoritam je algoritam za grupisanje u analitici podataka koji se primenjuje za razdvajanje skupa podataka u K grupa na osnovu sličnosti. K-means je dobro poznata tehnika particionisanja grupisanja. Ipak, ovaj pristup ima problem praznine klastera (Jain & Dubes,1988). Particiono ili neugnežđeno klasterisanje je podela skupa podataka na klastere koji sadrže slične podatke. Klastere generalno predstavlja centroid tj. centar klastera, odnosno podaci koji sumiraju opis podataka koji se nalaze u tom klasteru. Definicija centroida klastera zavisi od vrste podataka koje proučavamo; na primer, ako imamo dati skup realnih brojeva, onda će njihova težišta biti aritmetička sredina datog skupa. Ako je broj klastera veliki, centri se mogu dalje grupirati da bi se stvorila hijerarhija unutar skupa. U nastavku će biti analiziran

najpopularniji algoritam za grupisanje particija K-means algoritma, osim njega popularni su (Brin et al., 1997) i K-medoidi, fuzzi C-srednje vrednosti i algoritam maksimizacije očekivanja. Metoda particionog klasterovanja pokušava da maksimizuje sličnosti unutar klastera i minimizuje razlike između klastera.

Grupisanje K-means uveo je 1967. Džejms Mekvin. Ovaj algoritam koji se često koristi za grupisanje podataka je i danas popularan zbog svoje jednostavnosti i efikasnosti. Ovaj iterativni algoritam se naziva ciljno orijentisanim algoritmom i njegov cilj je da podeli podatke u K klastera. Osim K klastera, algoritam takođe daje i njihove predstavnike, pri čemu se minimizuje veličina funkcije cilja J.

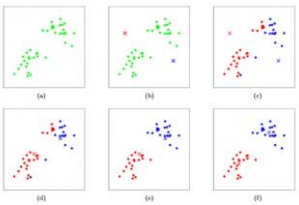
$$J(C) = \sum_{j=1}^{n} \sum_{k=1}^{K} u_{kj} \times ||x_j - c_k||^2$$

Težište klastera k se obeležava sa ck, a kj označava j-ti podatak (Jain & Dubes,1988). Početni centri klastera su nasumično odabrani podaci iz datog skupa Ks, a početna particija se formira po principu minimalnih rastojanja. U sledećim koracima algoritma težišta klastera se ažuriraju i postaju aritmetička sredina elemenata klastera. Ponavljamo proceduru particionisanja (ovaj korak se često naziva spajanjem) i ažuriranja centroida (tzv. korekcija) sve dok se ne dogodi jedan od sledećih uslova (Snajder & Bašić, 2015):

a. težišta klastera (centroidi) ostaju nepromenjena tokom iteracija,
b. vrednost funkcije cilja J postane manja od unapred definisane tolerancije,
c. dostignut je maksimalni broj iteracija koji je unapred predviđen.

Algoritam radi tako da se najprije odabere broj K grupa, a zatim se nasumično odabere K tačaka (centralnih tačaka) koje predstavljaju početne tačke grupa. Zatim, za svaku tačku iz skupa podataka se odredi kojoj centralnoj tački je najbliža, i ta tačka se doda u odgovarajuću grupu. Nakon toga se centralne tačke pomeraju tako da su srednje vrednosti tačaka u odgovarajućim grupama što bliže centralnoj tački. Ponavlja se proces sve dok se centralne tačke ne stabilizuju, tj. dok se ne postigne stabilna grupa (slika 1.). Konačan rezultat algoritma je podela skupa podataka u K grupa, gdje su tačke u svakoj grupi slične po nekom kriterijumu, obično euklidskoj udaljenosti (Snajder & Bašić, 2015).

Slika 1
ALGORITAM K–MEANS KROZ ITERACIJE



Izvor: . Snajder, B. Dalbelo Bašić, (2015) *Strojno učenje*, Skripta, Preuzeto sa sajta: https:// www.fer.unizg.hr

Postoje neke prednosti i nedostaci (Hahsler, 2015) korišćenja metode particionog klasterovanja. Prednosti su da se može nositi sa velikim skupovima podataka i da je brži od hierarhijske metode klasterovanja. Osim toga, zbog toga što se svaka tačka podataka samo stavlja u jedan klaster, interpretacija rezultata je lakša. Međutim, nedostaci metode particionog klasterovanja su što rezultati mogu zavisiti od početnih tačaka (centroida) koji su slučajno izabrani. Takođe, ako se skup podataka ne deli dobro u različite klastera, rezultati klasterovanja mogu biti manje tačni. Metoda particionog klasterovanja može se primeniti u oblastima, kao što su analiza tržišta, grupisanje kupaca i klasifikacija dokumenata.

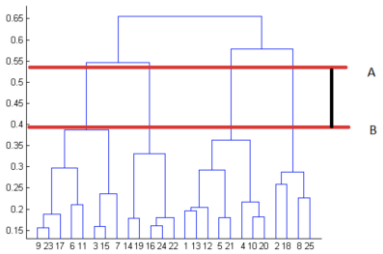
Hijerarhijsko grupisanje

Modeli povezivanja mogu imati jedan od dva pravca. U prvom metodu, oni počinju grupisanjem svih tačaka podataka u različite klastere, a zatim ih agregiraju kako rastojanje postaje sve manje. Drugi metod klasifikuje sve tačke podataka u jedan klaster pre nego što ih podeli kako rastojanje raste. Izbor funkcije udaljenosti je takođe proizvoljan. Iako su ovi modeli prilično jednostavni za razumevanje, oni nisu dovoljno skalabilni za rukovanje velikim skupovima podataka. Sledi da za razliku od particionog klasterisanja, hijerarhijsko klasterisanje rezultira hijerarhijom klastera. Primenjuju se dve grupe hijerarhijskog grupisanja: aglomerativno i divizijsko. Aglomerativno grupisanje „odozdo prema gore“ je proces spajanja klastera počevši od jednog podatka u svakom klasteru i spajajući postepeno parove klastera, dok se svi podaci ne spoje u jedan klaster. S druge strane, divizijsko grupisanje pristup „od vrha prema dole“, je proces koji kreće od pretpostavke da svi podaci pripadaju jednom klasteru, nakon čega se postepeno razdvajaju u slojevima hijerarhije. Ovaj proces spajanja i razdvajanja klastera se obično vrši na pohlepni način (Hájek et al., 2004).

Dok K-means algoritma ima teorijsku osnovu, hijerarhijsko grupisanje nema teorijsku osnovu i predstavlja heuristički postupak (Pei et al., 2001). Rezultat hijerarhijskog grupisanja je stablo sa granama koje predstavljaju formirane grupe. Na osnovu visine grananja, može se odrediti koliko su slične grupe međusobno. Hijerarhijsko grupisanje se često koristi u biologiji, genetici i društvenim naukama za formiranje klasifikacija i kategorizaciju (Scitovski & Sabo, 2020). Kao što ime kaže, hijerarhijsko grupisanje je algoritam koji stvara hijerarhiju klastera. Svaka od tačaka podataka dobija svoj sopstveni klaster na početku ovog procesa. Nakon toga, dva najbliža klastera se spajaju u jedan klaster. Ovaj metod na kraju prestaje kada ostane samo jedan klaster.

Hijerarhija klastera je prikazana dendrogramom. Dendrogram je drvo u kome listovi odgovaraju podacima, a horizontalne linije odgovaraju vezama na određenoj udaljenosti. Ovaj način prikaza grupisanja je zanimljiv jer se može preseći na bilo kojoj udaljenosti, što omogućava dobijanje klastera koji bi se dobili particionim grupisanjem na toj istoj udaljenosti (Tan et al., 2016). Dendrogram je vizuelni prikaz rezultata hijerarhijskog grupisanja.

Slika 2
OPTIMALAN IZBOR ZA BROJ KLASTERA



Izvor: . P.N. Tan, M. Steinbach, V. Kumar (2016) *Introduction to Data Mining*, Preuzeto sa sajta: https://www-users.cse.umn.edu/~kumar001/dmbook/sol.pdf

Dendrogram se može pročitati na sledeći način. Kao što je predstavljeno na slici 2. počinje se na dnu sa 25 tačaka podataka koje su podeljene u različite grupe. Nakon toga, dva najbliža klastera se kombinuju, ostavljajući samo jedan klaster na vrhu. Udaljenost između dva klastera u prostoru podataka je predstavljena visinom u dendrogramu na kojoj se spajaju. Gledajući dendrogram, može se odlučiti koliko će klastera najbolje predstavljati različite grupe. Broj vertikalnih linija u dendrogramu presečenih horizontalnom linijom koja može da pređe najveću vertikalnu udaljenost bez dodirivanja klastera je optimalan izbor za broj klastera. Crvena horizontalna linija dendrograma, koja predstavlja optimalan broj klastera u pomenutom slučaju, biće četiri pošto crvena horizontalna linija na dendrogramu ispod pokriva maksimalno vertikalno rastojanje AB.

Trebalo bi da se obrati pažnja na sledeća dva faktora o hijerarhijskom grupisanju (Snajder & Bašić, 2015):

- Za razvoj ovog algoritma korišćena je strategija odozdo prema gore. Takođe je moguće pratiti pristup odozgo prema dole počevši od svih tačaka podataka dodeljenih u istom klasteru i rekurzivno obavljajući podele sve dok se svakoj tački podataka ne dodeli poseban klaster.
- Izbor za spajanje dva klastera zasniva se na tome koliko su ti klasteri blisko povezani, počevši od svih tačaka podataka datih istom klasteru i rekurzivnog sprovođenja podela dok se svakoj tački podataka ne dodeli drugi klaster.

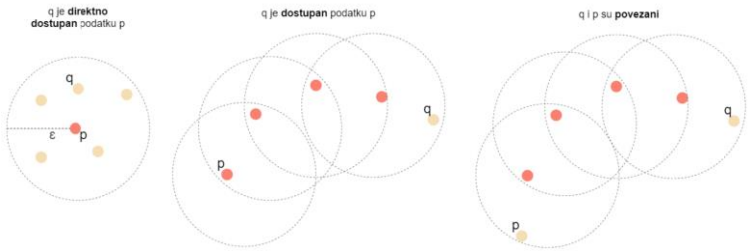
Blizina dva klastera može se odrediti korišćenjem različitih kriterijuma (Snajder & Bašić, 2015), uključujući:

- Euklidsko rastojanje: $||a-b||_2 = \sqrt{(\Sigma(a_i-b_i))}$
- Kvadrat Euklidskog rastojanja: $||a-b||_2^2 = \Sigma((a_i-b_i)^2)$
- Menhetn rastojanje: $||a-b||_1 = \Sigma|a_i-b_i|$
- Maksimalno rastojanje: $||a-b||_{INFINITY} = max_i|a_i-b_i|$
- Mahalanobisovo rastojanje: $\sqrt{((a-b)^T S^{-1} (-b))}$ {gde, s : kovarijansna matrica}

DBSCAN algoritam

DBSCAN (*eng. Density Based Spatial Clustering of Applications with Noise*) je algoritam u analitici podataka koji se koristi za grupisanje objekata u gusto naseljenim područjima. Ovaj algoritam je posebno namenjen za primenu u situacijama kada je potrebno grupisati objekte na osnovu gustine rasporeda, što ga čini korisnim za analizu podataka u različitim oblastima. Ideja je da se nastavi sa rastom datog klastera sve dok gustina u okruženju prelazi neki prag, tj. za svaku tačku podataka unutar datog klastera, radijus datog klastera mora da sadrži najmanje minimalni broj tačaka. DBSCAN algoritam (Xie et al., 2019) funkcioniše tako što se odredi početna tačka i epsilon-okolina (radijus okoline) koja određuje maksimalnu udaljenost objekata koji će biti grupisani u zajedničku grupu. Svaka tačka u epsilon-okolini se smatra dijelom iste grupe, a ako ne postoji dovoljno sličnih tačaka u okolini, ta tačka se smatra šumom (slika 3.).

Slika 3
PRAVA VREMENSKA DOSTUPNOST, PRISTUPAČNOST I MEĐUSOBNA POVEZANOST INFORMACIJA



Izvor: . Yiqun Xie et al. (2019) *A nondeterministic normalization based scan statistic (NN-scan) towards robust hotspot detection: a summary of results*. In SIAM Intl. Conf. on Data Mining (SDM'19).

DBSCAN algoritam koristi metriku udaljenosti za merenje udaljenosti između tačaka i definisanje susedstva oko svake tačke. Najčešće korišćena metrika udaljenosti je euklidska udaljenost, ali se mogu koristiti i druge metrike. Algoritam takođe ima dva glavna parametra (Han et al., 2012): epsilon (ε) i minimalan broj tačaka (MinPts). Osnovne formule koju koristi DBSCAN algoritam. Potrebno je za svaku tačku p u skupu podataka:

a. Identifikovati sve tačke q koje su udaljene ε od tačke p.

b. Ako je broj tačaka q veći ili jednak MinPts, tada je tačka p jezgro tačka i grupa se formira oko nje.

c. Ako p nije jezgro tačka, ali je udaljena ε od jezgra tačka, tada je tačka p granična tačka i dodaje se grupi najbliže jezgro tačke.

d. Ako tačka p nije udaljena ε od bilo koje jezgro tačke, tada je tačka p tačka buke.

Ponovljaju se gore navedeni koraci za sve tačke u skupu podataka dok sve tačke ne budu dodeljene grupi ili označene kao tačke buke. Formula se može modifikovati u

zavisnosti od specifične metrike udaljenosti i parametara algoritma koji se koriste. Glavna prednost (Tan et al., 2016) DBSCAN algoritma je mogućnost da identifikuje podskupove podataka koji su gusto raspoređeni, a koji mogu biti otkriveni samo na osnovu prostornog rasporeda podataka, bez pretpostavke o broju grupa ili njihovoj veličini. Rezultat DBSCAN algoritma su grupe objekata koji su gusto raspoređeni u prostoru, što ga čini korisnim u mnogim oblastima poput analize socijalnih mreža, urbanističkog planiranja, detekcije anomalija i drugima.

STING algoritam

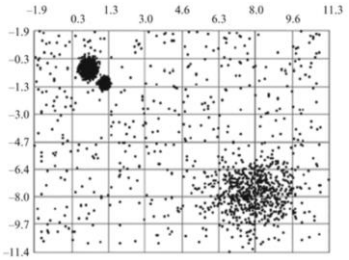
STING (*eng. Statistical information grid*) je popularan algoritam u ovoj metodi grupisanja. Grupisanje prostornih baza podataka je njena glavna upotreba. Grupisanje zasnovano na mreži se često koristi za identifikaciju klastera u ogromnim višedimenzionalnim prostorima, slično kao i klasterisanje zasnovano na gustini. Još jednom, na klastere se gleda kao na prepune oblasti (Bandyopadhyay & Saha, 2013).

Većina pristupa grupisanju ima linearnu vezu između veličine ulaznog skupa i njihove vremenske složenosti. Kapacitet za rukovanje ogromnim skupovima podataka je ključna prednost klasterisanja zasnovanog na mreži. Ključna razlika između klasterisanja zasnovanog na mreži i grupisanja zasnovanog na gustini je u tome što ovi algoritmi rade na okolnom prostoru, a ne na klasterizovanim podacima. Uobičajeni algoritmi zasnovani na mreži podataka često prate ove korake (Hipp et al., 2000):

Korak 1: Kreiranje mrežne strukture kao prvi korak. Da bi se to uradilo, prostor podataka se može podeliti na konačan broj mreža.
Korak 2: Izračunavanje gustine mreže sabiranjem svih podataka koje sadrži.
Korak 3: Sortiranje mreže po gustini.
Korak 4: Izračunavanje centroida klastera.
Korak 5: Prelazak na susednu mrežu.

Na početku se pravi mreža unutar prostora podataka. Hijerarhijska struktura služi kao reprezentacija ovih mreža. Koren hijerarhije je na nivou 1, a njeni potomci su na nivoima ispod. Ćelija na nivou I koja sadrži zbir svih njenih dečijih nivoa I + 1 regiona. Svaka ćelija u STING algoritmu ima četiri potomka. Kao rezultat, svako potomstvo zauzima petinu prostora roditeljske ćelije. STING je efikasan samo u dvodimenzionalnim situacijama (slika 4.).

Slika 4
PRIMER MREŽE PODATAKA



Izvor: . Hipp, J.; Güntzer, U.; Nakhaeizadeh, G. (2000) "*Algorithms for association rule mining -- a general survey and comparison*". ACM SIGKDD Explorations Newsletter 2: 58.

ZAKLJUČAK

Klasterovanje je proces grupisanja objekata koji pripadaju istoj klasi. Slični objekti su grupisani u jednom klasteru, dok su različiti objekti u drugom klasteru. Mnoge aplikacije koriste analizu klastera, uključujući istraživanje tržišta, identifikaciju obrazaca i analizu podataka.

Grupisanje K-means je brzo, pouzdano i lako razumljivo. Najbolji rezultati se dobijaju kada se skup podataka odvoji od drugog skupa podataka. Klasteri nisu hijerarhijskog oblika i ne preklapaju se. U slučaju da nijedna tačka podataka nije dodeljena klasteru, tokom faze dodeljivanja se formiraju prazni klasteri, što je problem. Predloženo je automatsko uklanjanje praznog klastera. Hijerarhijsko grupisanje koristi funkciju udaljenosti ili mere sličnosti u cilju pronalaženja klastera podataka koji su međusobno najsličniji. DBSCAN algoritam se zasniva na gustini, a STING algoritam na mreži i često se koristi za identifikaciju klastera u ogromnim višedimenzionalnim prostorima. Zbog dostupnosti ogromne količine informacija na mobilnim uređajima, napredna analiza podataka je logičan korak u domenu sveprisutnog računarstva. U mnogim industrijama, grupisanje se može koristiti za lociranje grupa korisnika, pacijenata, klijenata ili drugih objekata.

Rezultati ukazuju na veliko interesovanje i potrebu za daljim širenjem i usavršavanjem modela za grupisanje, jer se ovaj pristup koristi za istraživanje i pronalaženje skrivenih obrazaca i struktura u ogromnim skupovima podataka. Dokazani su ciljevi ovog rada, a tehnike klastiranja koje su navedene, zajedno sa drugim važnim algoritmima, predstavljaju ključnu osnovu u razvoju modernih sistema grupisanja podataka, koji se smatraju jednom od najčešćih tehnika analize podataka. Za buduća istraživanja planira se detaljnija analiza drugih algoritama grupisanja i unapređenje implementacije predloženih rešenja u oblasti grupisanja.

BIBLIOGRAFIJA

1. A. K. Jain, R. C. Dubes. (1988) Algorithms for Clustering Data, Prentice-Hall, Inc., USA.
2. Aggarwal, Charu C.; and Yu, Philip S. (1988) A new framework for itemset generation, in PODS 98, Symposium on Principles of Database Systems, Seattle, WA, USA, pages - 24.

3. Brin, Sergey; Motwani, Rajeev; Ullman, Jeffrey D.; and Tsur, Shalom. (1997) Dynamic itemset counting and implication rules for market basket data, in SIGMOD, Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 1997), Tucson, Arizona, USA, pp. 255-264.
4. Brown B, Sikes J, Willmott P. (2013) Bullish on digital: McKinsey Global Survey results, McKinsey. Quarterly, No. 12, pp. 1-8.
5. Dumbill E, (2012) "What is big data", An introduction to the big data landscape. Preuzeto sa sajta: www.mhsinformatics.org
6. Hájek, Petr; Feglar, Tomas; Rauch, Jan; and Coufal, David; (2004) The GUHA method, dana preprocessing and mining, Database Support for Data Mining Applications, Springer.
7. Hájek, Petr; Havel, Ivan; Chytil, Metoděj; (1966) The GUHA method of automatic hypotheses determination, Computing 1 293-308.
8. Higdon, D., Gattiker, J., Lawrence, E., Jackson, C., Tobis, M., Pratola, M., Habib, S., Heitmann, K., and Price, S. (2013) Computer model calibration using the ensemble kalman filter. In Technometrics, volume 55, pages 488–500. Taylor & Francis Group.
9. Hipp, J.; Güntzer, U.; Nakhaeizadeh, G. (2000) "Algorithms for association rule mining -- a general survey and comparison". ACM SIGKDD Explorations Newsletter 2: 58.
10. J. Han, M. Kamber, J. Pei, (2012) Data Mining: Concepts and Techniques, str. 443– 495.
11. J. Snajder, B. Dalbelo Bašić, (2015) Strojno učenje, Skripta, Preuzeto sa sajta: https:// www.fer.unizg.hr
12. Marz N, Warren J, (2014) Big data - Principles and best practices of scalable realtime data systems (Chapter 1).
13. Michael Hahsler. (2015) A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules. Preuzeto sa sajta: http://michael.hahsler.net/research/association_rules/mea sures.htm
14. Olfa Nasraoui, Chiheb-Eddine Ben N'Cir (2019) Clustering Methods for Big Data Analytics, Springer
15. P.N. Tan, M. Steinbach, V. Kumar (2016) Introduction to Data Mining, Preuzeto sa sajta: https://www-users.cse.umn.edu/~kumar001/dmbook/sol.pdf
16. Pei, Jian; Han, Jiawei; and Lakshmanan, Laks V. S. (2001) Mining frequent itemsets with convertible constraints, in Proceedings of the 17th International Conference on Data Engineering, April 2–6, Heidelberg, Germany, pages 433-442.
17. R. Scitovski, K. Sabo, (2020) Klaster analiza i prepoznavanje geometrijskih objekata, Sveučilište u Osijeku, Odjel za matematiku, Osijek
18. S. Bandyopadhyay, S. Saha, (2013) Unsupervised Classification, Springer-Verlag, Berlin Heidelberg.
19. Tan, Pang-Ning; Michael, Steinbach; Kumar, Vipin. (2005) "Chapter 6. Association Analysis: Basic Concepts and Algorithms" (PDF). Introduction to Data Mining. Addison-Wesley.
20. Yiqun Xie et al. (2019) A nondeterministic normalization based scan statistic (NN-scan) towards robust hotspot detection: a summary of results. In SIAM Intl. Conf. on Data Mining (SDM'19).

RESUME

The paper analyzes the importance of clustering in the analysis of large data sets and presents different clustering techniques. First, Big Data and clustering are defined, and then the most popular clustering techniques are described in detail: partitioning,

hierarchical clustering, density-based clustering, and data network-based clustering. Big Data refers to large amounts of data generated from various sources, such as social networks, sensors, e-commerce, etc. As the amount of data increases, so do the requirements for efficient and fast processing of that data. Cluster analysis is used as one of the methods for processing big data. Grouping similar objects into clusters helps in understanding data and uncovering hidden information. The most commonly used clustering techniques are described. K-means algorithm is used for partitioning, while dendrogram is used for hierarchical clustering. DBSCAN is used for density-based clustering and STING algorithm is used for network data. The paper highlights the importance of clustering in big data analysis, because the accuracy of big data can contribute to confidence in decision-making, and better estimates can help increase efficiency, reduce costs and risks.