SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Anto Čabraja

PARALELNI ALGORITMI ZA PROBLEM GRUPIRANJA PODATAKA

Diplomski rad

Voditelj rada:
prof. dr. sc. Goranka Nogo

Zagreb, srpanj 2014.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

0.1 Problem grupiranja podataka

0.2 Primjena

0.3 Pregled rada

Poglavlje 1

Modeliranje problema grupiranja

1.1 Osnovni pojmovi

Kako bi u daljnjem razmatranju bilo jednostavnije objašnjavati strukture i same implementacije algoritama potrebno je problem grupiranja reprezentirati osnovnim pojmovima. U nastavku ćemo formalno definirati sve komponente od kojih se problem grupiranja sastoji.

Definicija 1.1.1. ***Uzorak*** *je apstraktna struktura podataka koja reprezentira stvarne podatke s kojima raspolaže algoritam za grupiranje.*

Definicija 1.1.2. ***Svojstvo*** *je vrijednost ili struktura koja predstavlja jednu značajku danog podatka unutar uzorka.*

Definicija 1.1.3. ***Udaljenost*** *između uzoraka definiramo kao funkciju* $f : D- > \mathbb{R}$*, gdje je* D *skup svojstava danih uzoraka*

Definicija 1.1.4. *Za uzorke kažemo da su* ***blizu*** *jedan drugome ako je njihova udaljenost manja od unaprijed zadane veličine*

Definicija 1.1.5. ***Klaster*** *je skup uzoraka koji su u prostoru podataka blizu. Ako su uzorci identični onda je njihova udaljenost uvijek* 0

Definicija 1.1.6. ***Jednistveno grupiranje*** *je postupak grupiranja kada svaki uzorak pripada jednom i samo jednom klasteru.*

Definicija 1.1.7. ***Nejasno ili nejedinstveno grupiranje*** *je postupak grupiranja gdje jedan uzorak može biti u više klastera.*

Napomena 1.1.8. *U radu ćemo promatrati* ***jedinstveno grupiranje*** *tako da će sve daljnje definicje i modeliranja predpostavljati da želimo dobiti disjunktne klastere.*

1.2 Matematičko modeliranje problema

Definicija grupiranja podataka nije jedinstvena. U literati se na različite načine pokušava opisati ovaj postupak. Neki od pokušaja opisne definicije su:

1. *Grupiranje podataka je postupak otkrivanja homogenih*[1] *grupa uzoraka unutar skupa svih danih uzoraka.*

2. *Grupiranje podataka je postupak određivanja koji su uzorci slični te ih svrstati u isti klaster.*

Za modelirali problem neće nam biti dovoljne opisne definicije. U ovom slučaju opisne definicje mogu poslužiti samo kao intuicija o ćemu se zapravo radi kada govorimo o grupiranju. U nastavku ćemo pomoću definiranih pojmova u poglavlju 1.1 matematički opisati problem grupiranja podataka.

Neka je $\mathbf{U} = \{U_1, U_2, \cdots, U_n\}$ skup od n uzoraka,te neka je $U_i = (s_1, s_2, \cdots, s_d)$ reprezentiran d-dimenzionalnim vektorom gdje s_i predstavlja jedno svojstvo. Ovako definiran $\mathbf{U}$ moguće je reprezentirati kao matricu $\mathbf{S}_{d\times n}$.Svaki stupac te matrice predstavlja jedan uzorak iz danog skupa $\mathbf{U}$.

$$\mathbf{S}_{d\times n} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & \cdots & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ s_{d,1} & s_{d,2} & \cdots & \cdots & s_{d,n} \end{pmatrix} \tag{1.1}$$

Iz definicje 1.1.6 te iz navedenog formalnog zapisa dajemo formalnu definiciju problema grupiranja.

Definicija 1.2.1. ***Skup od k klastera*** $\mathbf{C} = \{C_1, C_2, ..., C_k\}$ *je skup sa sljedećim svojstvima:*

- $C_i \neq \emptyset$
- $C_i \cap C_j = \emptyset$, $\forall i, j \; i \neq j$
- $\cup_{i=1}^{k} C_i = \mathbf{U}$

Napomena 1.2.2. *U terminima matrice* $\mathbf{S}$ *to znači da se svaki* C_i *zapravo sastoji od stupaca matrice* $\mathbf{S}$*.*

Definicija 1.2.3. ***Problem grupiranja*** *u skup od k klastera* $\mathbf{C}$ *je ekvivalentan problemu da* $\forall c, c' \in C_i$ *udaljenost od c do c′ je manja od udaljenosti c do bilo kojeg drugog* $c'' \in C_j$ $j \neq i$

[1] podaci koji se ne mogu smisleno separirati

Zapravo problem grupiranja je pronalazak najpogodnije particije za $\mathbf{C}$ u skupu svih mogućih particija. Prema napomeni 1.2.2 lako se zaključi da se problem grupiranja svodi na problem raspodjele n stupaca matrice $\mathbf{S}$ u k skupova. Uvedena matematička notacija za problem grupiranja omogućuje da postavimo model za rješavanje, koji je u kasnijem razmatranju pogodan za modeliranje i implementaciju.
Neka je $C = \{\mathbf{C}^1, \mathbf{C}^2, ..., \mathbf{C}^{N(n,k)}\}$ skup svih mogućih rješenja danog problema grupiranja, gdje je

$$N(n,k) = \frac{1}{k!}\sum_{i=1}^{k}(-1)^i\binom{k}{i}(k-i)^i \tag{1.2}$$

broj mogućih rješenja za raspodjelu n uzoraka u k klastera. Rješenje problema svodi se na optimizacijski problem

$$\underset{\mathbf{C}\in C}{optimiziraj}\ f(\mathbf{S}_{d\times n}, \mathbf{C}) \tag{1.3}$$

gdje je f funkcija dobrote rješenjea $\mathbf{C}$, Funkcija dobrote je funkcija koja mjeri kvalitetu rješenja za dani problem. Ovisno o tome kako je zadana može se gledati problem maksimizacije ili minimizacije. Gotovo uvijek njome se određuju jedan od dva važna parametra:

- koliko su blizu uzorci u danom klasteru
- koliko su blizu dva disjunktna klastera

Ako promatramo udaljenosti uzoraka unutar klastera, onda nam se problem optimizacije 1.3 svodi na problem minimizacije funkcije f, jer želimo postići što manju udaljenost uzoraka unutar jednog klastera. Međutim ako želimo postići da su nam klasteri međusobno disjunktni i da granica disjunkcije bude čvrsto definirana moramo promatrati udaljenosti među klasterima. U ovom slučaju potrebno je maksimizirati problem to jest tražiti za koju particju će vrijednost funkcije f biti najveća.
Konačno, sada znamo uzorke reprezentirati kao n-dimenzionalne vektore, također znamo definirati klaster kao skup vektora, te smo postavili model za određivanje kvalitete određenog klastera. Preostali posao je pronaći konkretnu funkciju f koja će na adekvatan način reprezentirati udaljenost između uzoraka.Vrlo je važno definirari od koji se komponenti uzorak sastoji. Osnovna podjela uzoraka je na *numeričke* i *kategoričke*. Numerički uzorak je onaj uzorak čije su sve vrijednosti numeričkog tipa, dok je kategorički onaj uzorak čije vrijednosti poprimaju vrijednosti nekih kategorija. Više o mogućnostima izgleda svojstava uzorka biti će rečeno u cijelini 1.4. U ovom trenutku za definiranje funkcije dobrote potrebno je samo imati u vidu da podaci ne moraju biti jednostavni, ali i dalje se od funkcije dobrote očekuje da na adekvatan način odredi udaljenost između dva uzorka.

Uzorak s numeričkim značajkama

Ukoliko su nam uzorci takvi da ih možemo predstaviti kao vektor numeričkih podataka tada ih možemo smjestiti u vektoriski prostor, te na njima upotrijebiti neku od standardnih vektorskih normi. Za problem grupiranja podataka u praksi najčešće se koristi Euklidska ili neka od p-normi [liter]. Euklidska norma daleko je najpopularnija i glavni je predstavnik normi koje mjere različitost među uzorcima. Drugim rječima za Euklidsku normu vrijedi da su uzorci više različiti što je vrijednost norme veća.
Neka su S_1 i S_2 dva uzorka sa n značajki, tada udaljenost d između dva uzorka u eklidskoj normi računamo kao:

$$d(S_1, S_2) = \sqrt{\sum_{i=1}^{n} (s_{i,1} - s_{i,2})^2} \tag{1.4}$$

gdje su $s_{i,1}$ i $s_{i,2}$ značajke u uzorcima S_1 i S_2. U kreiranju rješenja sa numeričkim podacima uvijek ćemo koristit Euklidsku normu s malim izmjenama ovisno o vrsti problema. Međutim u praksi se vrlo često pojavljuje i noram koja koristi svojstva kovarijacijske matrice uzoraka[liter].

$$d(S_1, S_2) = (S_1 - S_2)^T \Sigma^{-1} (S_1 - S_2) \tag{1.5}$$

S_1 i S_2 uzorci, Σ kovarijacijska matrica uzoraka. I za ovo normu također vrijedi da su podaci više različiti što je vrijednost d veća. Također postoji veza između Euklidske i norme s kovarijacijskom matricom što je detaljno objašnjeno u [liter]. Samo ćemo reći da ako je Σ dijagonalna matrica onda se pripadna udaljenost zove normalizirana Euklidska udaljenost.
Osim normi koje mjere razlike među podacima, postoje i norme koje mjere sličnot. Norme koje mjere sličnost često su baziranje na određivanju kuta između dva uzorka u vektorskom prostoru. U [liter] objašnjene su neke od popularnijih normi ovog tipa.

Uzorak s kategoričkim značajkama

Individualna usporedba dva uzorka s kategoričkim značajkama vrlo često nema važnost i jedina povratna informacija je koliko značajki u uzorcima ima jednaku vrijednost. Ako ipak želimo postići da usporedbom dvije kategorije možemo zaključiti koliko su značajke u uzorku slične, to jest blizu, moramo svakoj značajki pridružiti težine i definirati operaciju razlike. Ako bolje pogledamo tim postupkom zapravo smo kategorije pretvorili u numeričke vrijednosti sa posebno definiranom funkcijom razlike. Promotrimo sljdeći primjer.

Primjer 1.2.4. *Neka nam je na raspolaganju skup podataka o cvijeću, te neka nam je svaki cvijet zadan kao vektor s tri kategoričke komponente. Ounačimo sa S i S' dva cvijeta iz*

skupa zadanih uzoraka.

$$S = (crven, amerika, iglicasto)$$
$$S' = (plav, europa, iglicasto)$$

Lako vidimo da jedini zaključak kojeg iz ovih podataka možemo dobiti jest da se na prve dvije komponente razlikuju, dok je na trećoj vrijednost kategorije ista. Ako bi naprimjer htjeli kategorijama pridružiti vrijednost morali bi definirati što znači da su amerika i europa različite i koliko nam je to bitno svojstvo. Važnost svojstva ima ključnu ulogu i vrlo je teško empirijski procijeniti kakve numeričke podatke dodjeliti kategorijama.

U praksi nismo uvijek u mogućnosti predvidjeti sve moguće kategoričke vrijednosti, pa samim time ne možemo svakoj kategoriji pridružiti numeričku vrijednost. Često korištena i popularna metoda za određivanje pripadnosti određenoj klasi, kada se radi o kategoričkim vrijednostima, je traženje udjela pojavljivanja određene kategoričke vrijednosti u promatranom klasteru.U tom slučaju ne promatramo udaljenost dva uzorka nego koliko u klasteru ima kategorički vrijednosti sličnih onima koje obilježavaju novi uzorak. Drugim rječima promatramo sličnost uzorka sa klasterom. Neka je $\mathbf{C} = \{C_1, C_2, ...C_n\}$ skup od n klastera. Neka je $c = (c_1, c_2, ..., c_d)$ uzorak sa d značajki istog tipa kao i uzorci iz zadanog skupa klastera. Definiramo funkciju sličnosti s kao

$$s(c, C_x) = \sum_{i=1}^{d} \frac{find(c_i, C_x)}{|C_x|} \tag{1.6}$$

gdje je $C_x \in \mathbf{C}$, funkcija $find$ vraća broj pojavljivanje značajke c_i na i-tom mjestu u svakom uzorku iz skupu C_x, a $|C_x|$ označava broj uzoraka u skupu C_x. Važno je naglasiti da se promatra samo i-to mjesto u svim značajkama. Postoji mogućnost da se vrijednost značajke, to jest kategorija, na i-tom i j-tom mjestu u uzorku podudaraju. Na primjer ako imamo uzorak koji predstavlja automobile, te su dvije od značajki danog uzorka boja interijera i boja automobila. Očito dvije navedene značajke mogu poprimiti istu vrijednost, ali zapravo su potpuno disjunktne i kao takve ih treba promatrati. Konačno c pridružimo onom klasteru za koji funkcija s vrati najveću vrijednost.
Kada promatramo određeni skup podataka i kreiramo vektore uzoraka, u svakom trenutku znamo veličinu tog uzorka. Prilikom modeliranja problema sami određujemo skup značajki koji koristimo. Saznanje o broju značajki i njihovoj važnosti možemo upotrijebiti kako bi svakoj značajki odredili težinu. Naime, kako nisu sve značajke jednako bitne želimo postići da se utjecaj nekih značajki restringira, odnosno naglasi. S obzirom na navedeni zahtjev modificiramo funkciju s iz 1.6 te definiramo novu funkciju s_w.

$$s_w(c, C_x) = \sum_{i=0}^{d} w_i \frac{find(c_i, C_x)}{|C_x|} \tag{1.7}$$

Kao što vidimo svakoj značajki unutar uzorka pridružili smo težinu w_i. Važno svojstvo funkcije s_w je da omogućuje potpuno isključivanje nekih od značajki. Navedeno svojstvo u praksi često služi za empirijsko otkivanje nebitnih značajki. Otkrivanje i uklanjanje značajki čije postojanje ne pridonosi poboljšanju rješenja uvelike smanjuje prostornu složenost što će detaljno biti obrađeno u 1.4.

Uzorak s hibridnim značajkama

Priroda problema za koje pokušavamo rješiti problem grupiranja često je složene strukture te je nemoguće podatke reprezentirati uzorcima za značajkama samo jednog tipa. U takvom slučaju podatak se reprezentira za uzorkom čije značajke mogu biti numeričkog i kategoričkog tipa. Ovako definirana reprezentacija podataka povećava složenost i modeliranje funkcije dobrote, odnosno odluka o tome koliko su dva uzorka slična. Označimo sa $S = (s_1, s_2, ..., s_n)$ uzorak od n značajki numeričkog i kategoričkog tipa. Bez smanjenja općenitosti možemo predpostaviti da su na prvih l komponenti numeričke značajke, a na ostalih $n - l$ kategoričke. Funkciju udaljenosti d_h između dva uzorka S_1 i S_2 s hibridnim značajkama definiramo kao

$$d_h(S_1, S_2) = d(S_1(: l), S_2(: l)) + \frac{1}{s_w(S_1(l+1:), C_{S_2})} \tag{1.8}$$

gdje je d funkcija definirana u 1.4, s_w definirana u 1.7, a C_{S_2} označava klaster u kojemu se nalazi S_2. Važno je napomenuti da S_1 ne mora nužno biti u klasteru dok S_2 mora. Dakle ako dodajemo novi uzorak i želimo mu pronaći klaster to ćemo učiniti tako da ga stavimo kao prvi argument funkcije d_h. Kao što vidimo vrijednost funkcije s_w može biti jednaka nuli pa nam drugi sumand u definiciji funkcije d_h nije definiran. U praksi se eksperimentalno odredi kolika je važnost da uzorak nema niti jednu katekoričku značajku istu kao značajke unutar promatranog klastera, te se funkcija d_h dodefinira s obzirom na slučaj kada je $s_w = 0$. Neka je λ eksperimentalno određena vrijednost tada se potpunda definicija funkcije d_h može zapisati

$$d_h(S_1, S_2) = \begin{cases} d(S_1(: l), S_2(: l)) + \frac{1}{s_w(S_1(l+1:), C_{S_2})} & s_w \neq 0 \\ \\ d(S_1(: l), S_2(: l)) + \lambda & s_w = 0 \end{cases} \tag{1.9}$$

Ovako definitrana funkcija d_h je korektna. Naime, za sve parove uzoraka uvijek možemo odrediti vrijednost funkcije d_h jer su sve komponente dane funkcije dobo definirane. Funkcija d je zapravo standardna Euklidska norma pa je kao takva definirana za sve vektore. Funkcija s_w je kompozicija funkcije $find$, funkcije zbrajanja, množenja i kardinalnog broja. Sve funkcije su dobro definirane[liter] pa je i kompozicija dobro definiran. Primjerimo da se ovako definirana funkcija d_h, sa pravilno postavljenim paremetrom λ može koristit za sve slučajeve uzoraka: numeričke, kategoričke ili hibridne.

1.3 Metode razvoja algoritama za grupiranje

U postupku modeliranja rješenja za problem grupiranja koristi se više metoda. Metoda za rješavanje ima puno i znastvenici intenzivno rade na pronalasku novih i objašnjenju kvalitete postojećih metoda. Danas je ovo područje izuzetno cjenjeno i svako novo saznanje može dovesti do revolucionarnog napretka u rješavanju izuzetno teških problema. Ipak, dvije metode su se pokazale kao općenitije i mogu poslužiti kao predložak za rješavanje većeg broja problema. Konkretno radi se o *hijerarhijskom* i *particijskom* pristupu rješavanju problema. Među ostalim metodama koje se danas ističu kao jako korisne pokazale su se : evolucijske metode, metode particioniranja grafa, metode bazirane na neuronskim mrežama, te metode s gradinjentnim spustom[liter]. U ovom radu obradit ćemo navedene najpoznatije metode, a od ostalih metoda obradit ćemo evolucijske. Također pokazat ćemo kako se komponente iz evolucijskih metoda mogu kombinirati sa particijskim pristupom, te će ta hibridna kombinacija biti glavna tema u nastavku ovog rada.

Hijerarijsko grupiranje

1.4 Upravljanje podacima

Poglavlje 2

Meta-heuristički pristup problemu

2.1 Prirodom inspirirani algoritmi

2.2 Reprezentacija podataka

2.3 Analiza rezultata

Poglavlje 3

Poznati algoritmi i analiza

3.1 Alg 1

3.2 Alg 2

3.3 Alg 3

Poglavlje 4

Tehnike za paralelizaciju algoritama

4.1 Osnovni pojmovi MPI tehnologije

4.2 Topologije

4.3 Prednosti paralelizacije i cijena komunikacije

Poglavlje 5

Konstrukcija paralelnih algoritama za grupiranje

5.1 Algoritam 1 heurisika

Opis

Analiza

5.2 Algoritam 2 iterativno

Opis

Analiza

5.3 Algoritam 3 hibrid

Opis

Analiza

Poglavlje 6

Ostale moderne metode

6.1 Programiranje na grafičkim karticama

6.2 MapReduce metoda

Bibliografija

Sažetak

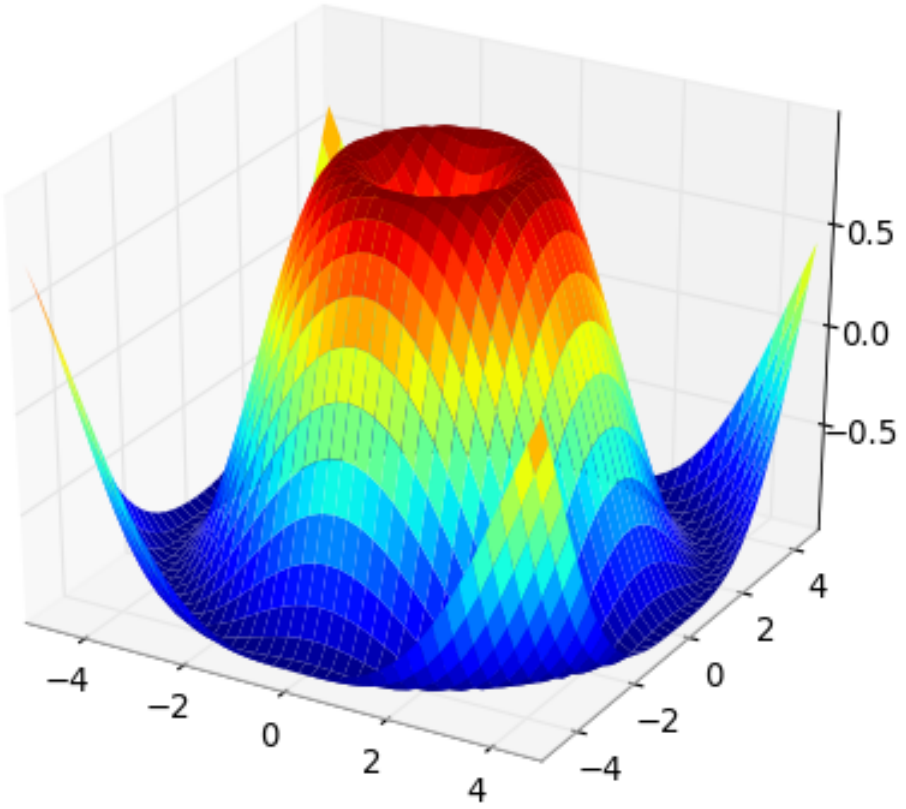
Ukratko ...

Summary

In this ...

Životopis

Na slici 1.1.7 se nalazi 3D graf neke funkcije.



Slika 6.1: Druga slika

kao i jedna vrlo kompliciran formula koja slijedi iz (**??**)

$$\sum_{i=1}^{\infty} A_{x_{1}} \times A_{\alpha_{2}} \oslash \iint_{\Omega} x^{2} \ddagger \limsup_{n \in \mathbb{N}} \frac{\alpha + \theta + \gamma}{n^{\omega}} \text{ je u stvari } \bigcup_{r \in \mathbb{Q}}^{\left(+ \right)} \overline{\Xi_{i} \underset{\substack{j \in \mathbb{C} \\ j \ni i \mathbb{Q}}}{\Theta} \Upsilon^{k^{j}} \underset{*}{\Psi} \hbar |_{\{\alpha\}}}.$$