

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

Sadržaj	iii
0.1 Problem grupiranja podataka	1
0.2 Primjena	1
0.3 Pregled rada	1
1 Modeliranje problema grupiranja	3
1.1 Osnovni pojmovi	3
1.2 Matematičko modeliranje problema	4
1.3 Metode razvoja algoritama za grupiranje	6
1.4 Upravljanje podacima	6
2 Metaheuristike	7
2.1 Prirodom inspirirani algoritmi	7
2.2 Reprezentacija podataka	7
2.3 Analiza rezultata	7
3 Poznati algoritmi i analiza	9
3.1 Alg 1	9
3.2 Alg 2	9
3.3 Alg 3	9
4 Tehnike za paralelizaciju algoritama	11
4.1 Osnovni pojmovi MPI tehnologije	11
4.2 Topologija	11
4.3 Prednosti paralelizacije i cijena komunikacije	11
5 Konstrukcija paralelnih algoritama za grupiranje	13
5.1 Algoritam 1 heurisika	13
5.2 Algoritam 2 iterativno	13
5.3 Algoritam 3 hibrid	13

6 Ostale moderne metode	15
6.1 Programiranje na grafičkim karticama	15
6.2 MapReduce metoda	15
Bibliografija	17

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 $f : D \rightarrow \mathbb{R}$, 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. *Jedinstveno 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 definicije i modeliranja pretpostavljati da želimo dobiti disjunktne klastere.

1.2 Matematičko modeliranje problema

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

1. *Grupiranje podataka je postupak otkrivanja homogenih¹ grupa uzoraka unutar skupa svih danih uzoraka.*
2. *Grupiranje podataka je postupak određivanja koji su uzorci slični te ih svrstati u isti klaster.*

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

Neka je $\mathbf{U} = \{U_1, U_2, \dots, U_n\}$ skup od n uzoraka, te neka je $U_i = (s_1, s_2, \dots, s_d)$ 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} .

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

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

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

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

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 $c'' \in C_j$ $j \neq i$*

¹podaci 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 $C = \{\mathbf{C}^1, \mathbf{C}^2, \dots, \mathbf{C}^{N(n,k)}\}$ skup svih mogućih rješenja danog problema grupiranja, gdje je

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

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

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

gdje je f funkcija dobrote rješenja \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 particiju 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 definirati 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 svojstvima

Ukoliko su nam uzorci takvi da ih možemo predstaviti kao vektor numeričkih podataka tada ih možemo smjestiti u vektorski 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 euklidskoj normi računamo kao:

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

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 koristiti Euklidsku normu s malim izmjenama ovisno o vrsti problema. Međutim u praksi se vrlo često pojavljuje i norma koja koristi svojstva kovarijacijske matrice uzoraka[liter].

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

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 različitost podataka, postoje i norme koje mjere sličnost. Norme koje mjere sličnost često su bazirane na određivanju kuta između dva uzorka u vektorskom prostoru. U [liter] objašnjene su neke od popularnijih normi ovog tipa.

1.3 Metode razvoja algoritama za grupiranje

A

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

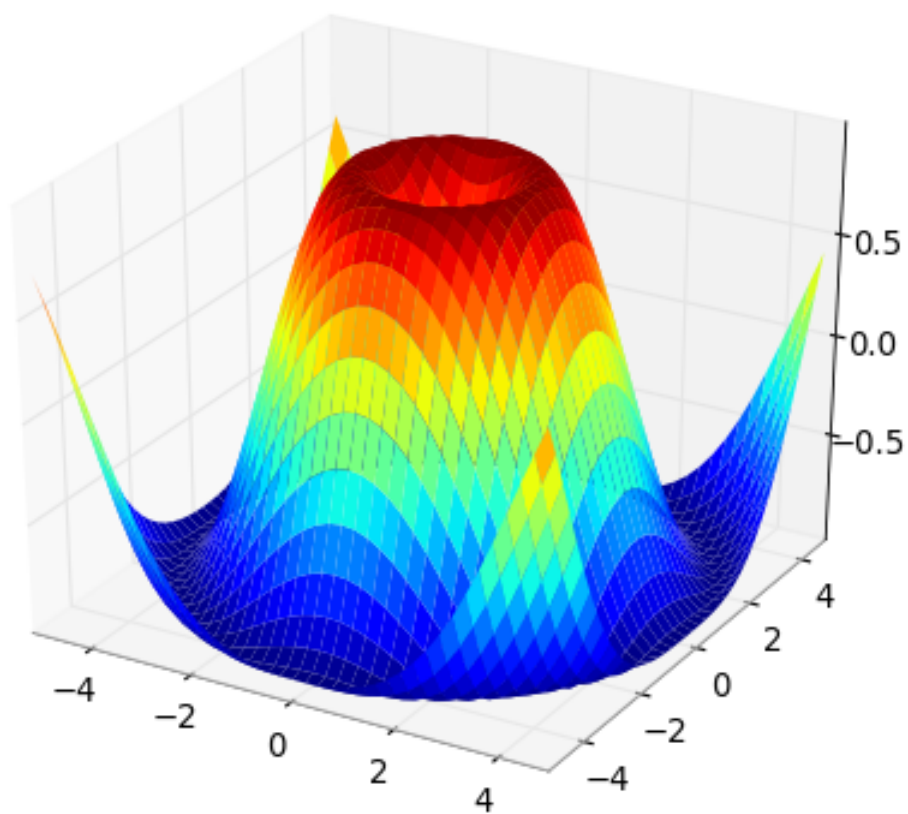
Ukratko ...

Summary

In this ...

Životopis

Na slici 1.1.7 se nalazi 3D graf neke funkcije.



Slika 6.1: Druga slika

kao i jedna vrlo komplicirana formula koja slijedi iz (??)

$$\sum_{i=1}^\infty A_{x_1}\times A_{\alpha_2}\oslash\iint_\Omega x^2\ddagger\limsup_{n\in\mathbb{N}}\frac{\alpha+\theta+\gamma}{n^\omega}\text{ je u stvari }\bigcup_{r\in\mathbb{Q}}\overline{\Xi_i\ominus_{\substack{j\in\mathbb{C}\\j\ni i\mathbb{Q}}}\Upsilon^{kj}_*\Psi\hbar|_{\{\alpha\}}}.$$