Neizrazito, evolucijsko i neuroračunarstvo
Neizrazito grupiranje

dr.sc. Marko Čupić

Fakultet elektrotehnike i računarstva
Sveučilište u Zagrebu

23. siječnja 2014.

Grupiranje

- Pretpostavimo da imamo skup od n uzoraka u r dimenzijskom prostoru: $\mathbf{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ gdje je $\vec{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,r})$.
- Pretpostavimo da se u tom prostoru uzorci grupiraju u c razreda.
- Zadaća *klasičnog* grupiranja je rastaviti skup uzoraka $\mathbf{X}$ na uniju c disjunktnih podskupova A_i, odnosno želimo da vrijedi:

$$\bigcup_{i=1}^{c} A_i = \mathbf{X}$$

$$A_i \bigcap A_j = \emptyset \;\; \forall i \neq j$$

$$\emptyset \subset A_i \subset \mathbf{X} \;\; \forall i.$$

- Vrijedi: $2 \leq c < n$

Grupiranje

- Jedan "algoritam" grupiranja već smo upoznali $\Rightarrow$ kod samoorganizirajućih neuronskih mreža:
  - mreža INSTAR radila je grupiranje podataka odnosno učila pozicije reprezentanata uzoraka iz ulaznog prostora
  - Kohonenova mreža SOM radila je grupiranje uz dodatno uvođenje topološke strukture između pojedinih razreda
- U literaturi je moguće pronaći više "klasičnih" algoritama grupiranja
  - spomenimo kao najpoznatiji $c - means$ (Bezdek, 1981)

Rezultat klasičnog grupiranja

Rezultat klasičnog grupiranja moguće je predstaviti na više načina; često se koristi matrični zapis.

- Neka je $\chi_{i,j}$ indikatorska funkcija (domena $\{0,1\}$) koja govori pripada li i-ti uzorak j-tom razredu.
- Matrica $\mathbf{M} = [\chi_{i,j}]$ je matrica koja ima onoliko redaka koliko ima uzoraka i onoliko stupaca koliko ima razreda i predstavlja rezultat grupiranja.
- U svakom retku matrice $\mathbf{M}$ samo je jedan element vrijednosti 1 dok su svi ostali vrijednosti 0.

Rezultat klasičnog grupiranja

Evo jednostavnog primjera: imamo 10 uzoraka i tri grupe.

Uzorak	Grupa 1	Grupa 2	Grupa 3
Uzorak 1	0	1	0
Uzorak 2	0	1	0
Uzorak 3	1	0	0
Uzorak 4	1	0	0
Uzorak 5	0	1	0
Uzorak 6	0	0	1
Uzorak 7	0	0	1
Uzorak 8	0	0	1
Uzorak 9	0	0	1
Uzorak 10	1	0	0

Zadatak

Kod neizrazitog grupiranja uklanjaju se oštre granice: dozvoljeno je da uzorak svakom od razreda pripada u određenoj mjeri.

- Neka je $\mu_{i,j}$ funkcija pripadnosti (domena $[0,1] \subset \mathbb{R}$) koja govori u kojoj mjeri i-ti uzorak pripada j-tom razredu.
- Matrica $\mathbf{M} = [\mu_{i,j}]$ tada je matrica koja ima onoliko redaka koliko ima uzoraka i onoliko stupaca koliko ima razreda i predstavlja rezultat grupiranja.
- U svakom retku matrice $\mathbf{M}$ više elemenata može imati vrijednost različitu od 0. Međutim, uobičajeno se postavlja zahtjev da njihova suma mora biti 1 odnosno "jedinična" pripadnost dijeli se između više razreda:

$$\sum_{j=1}^{c} \mu_{i,j} = 1.$$

Rezultat klasičnog grupiranja

Evo jednostavnog primjera: imamo 10 uzoraka i tri grupe.

Uzorak	Grupa 1	Grupa 2	Grupa 3
Uzorak 1	0.1	0.9	0
Uzorak 2	0.1	0.8	0.1
Uzorak 3	1	0	0
Uzorak 4	0.9	0	0.1
Uzorak 5	0.2	0.6	0.2
Uzorak 6	0.1	0.3	0.6
Uzorak 7	0.2	0	0.8
Uzorak 8	0	0	1
Uzorak 9	0	0	1
Uzorak 10	0.9	0	0.1

Ideja

Algoritam *fuzzy c-means* grupiranje ne radi direktno već za svaku grupu definira njezin centar.

- označimo centar i-te grupe oznakom $\vec{v}_i$
- $\vec{v}_i$ je točka u r-dimenzijskom prostoru baš kao što su to i ulazni uzorci skupa $\mathbf{X}$

Grupiranje se obavlja temeljem udaljenosti između promatranih uzoraka i centara grupa. Definira se funkcija cilja:

$$J(M, v) = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{i,j}^{m} \cdot (d_{i,j})^2$$

gdje je $\mu_{i,j}$ mjera kojom uzorak $\vec{x}_i$ pripada grupi j čiji je centar $\vec{v}_j$, m je parametar koji određuje "jakost" neizrazitosti grupiranja ($m \geq 1$) a $d_{i,j}$ je udaljenost između uzorka $\vec{x}_i$ i centra $\vec{v}_j$.

Ideja

- Zadaća neizrazitog grupiranja možemo definirati na sljedeći način: uz zadan način izračuna udaljenosti (primjerice, Euklidska udaljenost) pronaći svih c centara uz koje funkcija $J(M, v)$ poprima minimalnu vrijednost.

Neka se udaljenost $d_{i,j}$ računa na sljedeći način:

$$d_{i,j} = \sqrt{\sum_{k=1}^{r}(x_{i,k} - v_{j,k})^2}$$

Ideja

Znamo li pozicije centara grupa, mjeru pripadnosti svakog uzorka svakoj od grupa izračunat ćemo na temelju blizine uzorka svakom od centara:

$$\mu_{i,j} = \frac{\left(\frac{1}{d_{i,j}}\right)^{\frac{2}{m-1}}}{\sum_{k=1}^{c}\left(\frac{1}{d_{i,k}}\right)^{\frac{2}{m-1}}}$$

Izraz zapravo funkciju udaljenosti $d_{i,j}$ transformira u funkciju blizine $s_{i,j} = \left(\frac{1}{d_{i,j}}\right)^{\frac{2}{m-1}}$ i potom kao mjeru pripadnosti uzorka i grupi j definira omjer bliskosti uzorka i centru grupe j i sume bliskosti uzorka i centrima svih grupa.

Ideja

Na taj način automatski imamo zadovoljeno i:

$$\sum_{k=1}^{c} \mu_{i,k} = 1$$

Izraz za $\mu_{i,j}$ češće pišemo na sljedeći način:

$$\mu_{i,j} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{i,j}}{d_{i,k}} \right)^{\frac{2}{m-1}}}$$

Prilikom izračuna $\mu_{i,j}$, moguće je da se i-ti uzorak podudari s jednim ili više centara (tj. da je $d_{i,k} = 0$). Ako to nije slučaj, mjera pripadnosti se računa prema danom izrazu. Ako se pak i-ti uzorak podudari s l od c centara, mjera pripadnosti uzorka i grupama čiji su to centri ručno se postavi na $\frac{1}{l}$ dok se mjera pripadnosti tog uzorka preostalim grupama postavi na 0.

Ideja

Jednom kada su izračunate mjere pripadnosti, moguće je izračunati nove pozicije centara grupa tako da bolje aproksimiraju svoju grupu.

- Kod klasičnog grupiranja, centar bi bio aritmetička sredina svih uzoraka koji pripadaju grupi.
- Kod neizrazitog grupiranja, centar se računa kao težinska suma gdje su težine jednake mjeri pripadnosti:

$$\vec{v}_j = \frac{\sum_{i=1}^{n} \mu_{i,j}^m \cdot \vec{x}_i}{\sum_{i=1}^{n} \mu_{i,j}^m} \quad \forall j \in \{1, \cdots, c\}$$

odnosno raspisano po komponentama:

$$v_{j,k} = \frac{\sum_{i=1}^{n} \mu_{i,j}^m \cdot x_{i,k}}{\sum_{i=1}^{n} \mu_{i,j}^m} \quad \forall j \in \{1, \cdots, c\}, k \in \{1, \cdots, r\}$$

Algoritam

Pseudokod algoritma *fuzzy c-means* tada je dan u nastavku.

1. Odaberi $1 \leq m$, željeni broj grupa $1 < c < n$ te kriterij zaustavljanja. Kao početne centre $\vec{v}_i$ odaberi slučajno neke od uzoraka iz skupa koji se grupira.
2. Ponavljaj
   1. Izračunaj udaljenosti svakog od uzoraka do svakog od centara: $d_{i,j}$.
   2. Izračunaj mjere pripadnosti svakog od uzoraka svakoj od grupa: $\mu_{i,j}$.
   3. Izračunaj nove pozicije centara grupa.
   4. Ako je promjena u pozicijama centara dovoljno mala, prekini postupak.

Dodatna razmatranja

- za $m = 1$ mjere pripadnosti će težiti prema 0 ili 1 što algoritam pretvara u klasično grupiranje.
- kako m raste, to će se mjere pripadnosti smanjivati odnosno više distribuirati po grupama (raste neizrazitost).
- prethodno dani algoritam rješava optimizacijski problem traženja minimuma $J(M, v)$: međutim, rezultat grupiranja ovisi o početno odabranim centrima i može često zapeti u lokalnom optimumu (loše grupiranje)
- umjesto danog algoritma problem se može napasti algoritmima evolucijskog računanja (bilo za odabir dobrih početnih centara, bilo za traženje konačnih vrijednosti centara).

Dodatna razmatranja

- kako ocijeniti koliko nam centara doista treba ako to ne znamo unaprijed?
  - postoje različite ocjene koje se mogu koristiti
- kriterij zaustavljanja?
  - kada je suma euklidskih udaljenosti centra prije korigiranja i centra nakon korigiranja manja od zadane vrijednosti
  - ako vektore centara posložimo u matricu, možemo računati neku od matričnih normi razlike matrice centara prije i nakon ažuriranja, i tražiti da je ta norma manja od neke zadane
  - fiksan broj iteracija (loš pristup)
- postoje modifikacije načina izračuna udaljenosti koji nije simetričan po svim dimenzijama (što rezultira sferom) nego nekim dimenzijama daje veći utjecaj (hiperelipsoid): algoritam se bolje može prilagoditi podatcima različitog oblika

Riješeni primjer

Na Ferku u repozitoriju nalazi se `c-means-primjer.txt`.

- Sadrži detaljno riješen primjer neizrazitog grupiranja algoritmom *fuzzy c-means*.
- 12 uzoraka
- 3 razreda
- $m = 2$
- Uvjet zaustavljanja opisan u dokumentu.

Interesantan primjer

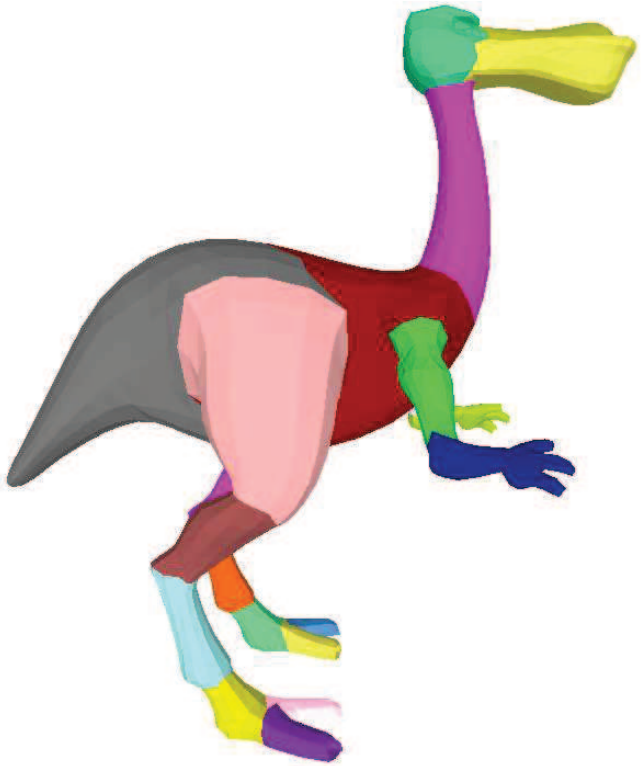
Neizrazito grupiranje često se koristi pri obradi slike. Prikazat ćemo primjer opisan u radu:
Katz, Sagi and Tal, Ayellet. Hierarchical Mesh Decomposition Using Fuzzy Clustering and Cuts. ACM SIGGRAPH 2003 Papers, p. 954–961, ACM, New York, NY, USA, 2003.
`http://webee.technion.ac.il/~ayellet/Ps/0325_ayt.pdf`

Interesantan primjer

- Za 3D tijelo dano je oplošje koje je modelirano mrežom poligona (sjetite se Interaktivne računalne grafike – opis pogodan za vizualizaciju i različite modele osvjetljavanja).
- To znači da za tijelo imamo stotine (ili tisuće) poligona.
- Ideja je automatski pronaći koji sve poligoni pripadaju istoj komponenti tijela (primjerice, kod čovjeka: lijeva ruka, desna ruka, glava, ...).

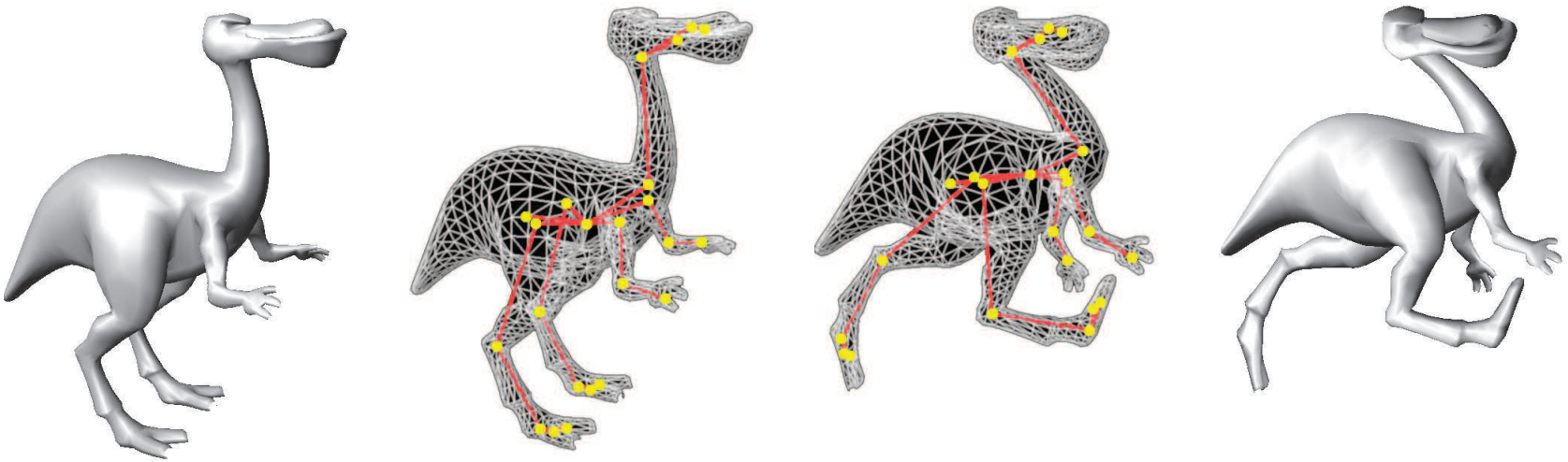
Interesantan primjer

Ovo želimo dobiti:



Interesantan primjer

Ovo je motivacija:



(a) object (b) skeleton (c) deformed skeleton (d) deformed object

Priprema

- tijelo je opisano mrežom poligona
- želimo pronaći dekompoziciju u "smislene" komponente
- za svaka dva susjedna poligona (dijele brid!) definira se kutna i geodezijska udaljenost (ideja: što je ona veća, manja je vjerojatnost da oba poligona pripadaju istoj komponenti; točan izračun sada nije bitan)
- gradi se dualni graf (vrhovi su poligoni, bridovi postoje između povezanih poligona)
- u njemu se definiraju težine bridova proporcionalne udaljenosti poligona
- udaljenost proizvoljna dva poligona tada se računa kao duljina najkraćeg puta u dualnom grafu

Binarni slučaj

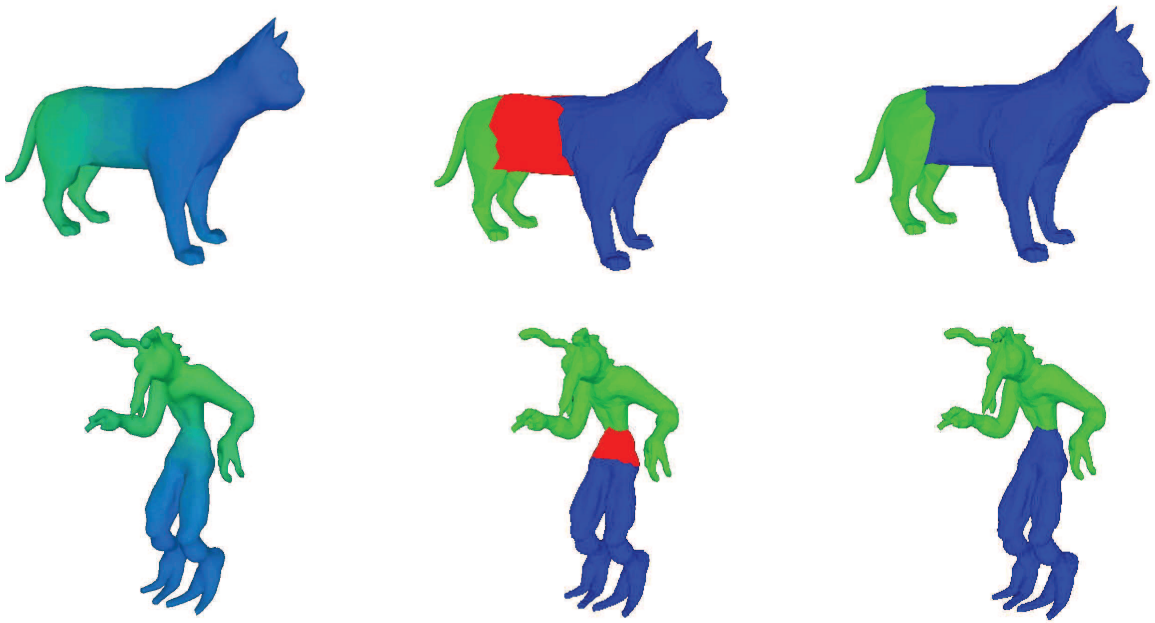
Pretpostavimo sada da postoje samo dvije komponente u tijelu.

Zadaća: treba napraviti neizrazito grupiranje kako bi se odredilo koji poligoni pripadaju kojoj od te dvije komponente.

- neki poligoni "jako" će pripada ili jednoj ili drugoj komponenti
- u području gdje se komponente spajaju, manje će biti jasna pripadnost: očita potreba za neizrazitošću
- dobit ćemo neizrazite granice koje algoritam naknadno obrađuje (to nas dalje ne zanima)

Binarni slučaj

Pretpostavimo sada da postoje samo dvije komponente u tijelu.



(a) probabilities (b) fuzzy decomposition (c) decomposition

Crveno su prikazani poligoni koji pripadaju i jednoj i drugoj komponenti.

Opći slučaj

Pretpostavimo sada da postoji k komponenata u tijelu.

Zadaća: treba napraviti neizrazito grupiranje kako bi se odredilo koji poligoni pripadaju kojoj od tih k-komponenata.