Wybrane aspekty wykrywania wspólnot w grafie

Mieczysław Kłopotek
Instytut Podstaw Informatyki
Polskiej Akademii Nauk

VI Spotkania Polskiej Grupy Badawczej Systemów Uczących Się
Częstochowa, 14 kwietnia 2016

Wybrane aspekty wykrywania wspólnot w grafie

Mieczysław Kłopotek
Instytut Podstaw Informatyki
Polskiej Akademii Nauk

VI Spotkania Polskiej Grupy Badawczej Systemów Uczących Się
Częstochowa, 14 kwietnia 2016

Agenda

1. Czym są wspólnoty - wspólnoty i grafy
2. Wymagania na funkcję oceny jakości skupień jako formalne wymaganie wobec algorytmów
3. Jakość skupień w praktyce
4. Algorytmy
5. Podsumowanie

Agenda

Wspólnota

Wspólnota

Skupienie w grafie sieci społecznej, generalnie w grafie.

Graf

$G = (V, E, L, l)$
V - zbiór wierzchołków, $E \subset V \times V$, - zbiór krawędzi
$L \subset \mathbb{R}$ - zbiór etykiet, $l : E \rightarrow L$ - funkcja etykietująca krawędzie
Zakładamy grafy nieskierowane, czyli $(u, v) \in E$ implikuje $(v, u) \in E$

Analiza skupień (dla grafów)

Taki podział zbioru wierzchołków na (zwykle) rozłączne podzbiory, by elementy tego samego zbioru były "podobne", a różnych zbiorów "niepodobne" .
Formalnie algorytm analizy skupień $\mathcal{A}$ to funkcja $\mathcal{A} : G \rightarrow 2^{2^V}$ taka, że jeśli $\mathcal{A}(G) = \mathcal{C}$, to $\cup_{C \in \mathcal{C}} = V$ i $C_1 \cap C_2$ dla $C_1, C_2 \in \mathcal{C}, C_1 \neq C_2$.

Co reprezentuje graf

- Grafy (w tym empiryczne)
- Sieci komunikacyjne, transportowe
- Sieci społeczne
- Rozmaitości niskowymiarowe w przestrzeniach wysokowymiarowych

Graf za każdym razem jest abstrakcją – wymaga zdefiniowania odwzorowania

Co reprezentują etykiety krawędzi grafu

- Odległości
- "Pseudoodległości"
- Niepodobieństwa
- Podobieństwa

Algorytmy analizy skupień w grafach zwykle operują na podobieństwie – potrzebna konwersja

Agenda

Wymagania na funkcję oceny jakości skupień, wg vanLaarhoven:2014

Funkcja oceny jakości podziału

Funkcja Q oceny jakości podziału grafu G odwzorowuje jego podziały $\mathcal{C}$ na zbiór liczb rzeczywistych $\mathbb{R}$.

Własności funkcji oceny jakości podziału:

- jeśli f jerst funkcją izomorficznie przekszrtałcającą zbiór V w sam siebie, to $Q(G, \mathcal{C}) = Q(f(G), f(\mathcal{C}))$. (niezmienniczość ze wzgledu na permutację)
- jeśli graf G^α to graf G, w którym wszystkie wagi krawędzi pomnożono przez stałą $\alpha > 0$, to jeśli dla podziałów $\mathcal{C}_1, \mathcal{C}_2$ zachodzi $Q(G, \mathcal{C}_1) \leq Q(G, \mathcal{C}_2)$, to $Q(G^\alpha, \mathcal{C}_1) \leq Q(G^\alpha, \mathcal{C}_2)$ (niezmienniczość ze wzgledu na skalę)
- Dla dowolnego zbioru V i jego nietrywialnego podziału $\mathcal{C}^*$ istnieje taki graf $G = (V, E, L, i)$, że $Q(G, \mathcal{C}^*) = \arg\max_{\mathcal{C}} Q(G, \mathcal{C})$ (bogactwo)

CDN.

Wymagania na funkcję oceny jakości skupień, wg vanLaarhoven:2014

Własności funkcji oceny jakości podziału, c.d.:

- małe zmiany w wartości funkcji podobieństwa l prowadzą do małych zmian wartości funkcji Q (ciągłość)
- Dla dwóch grafów $G_1 = (V, E, L, l_1)$ i $G_2 = (V, E, L, l_2)$ oraz podziału $\mathcal{C}$ takich, że $l_1((u,v)) \leq l_2((u,v))$ dla u, v z tego samego skupienia i $l_1((u,v)) \geq l_2((u,v))$ dla u, v z różnych skupień $Q(G_1, \mathcal{C}) \leq Q(G_2, \mathcal{C})$ (spójna poprawa)
- Dla trzech rozłącznych zbiorów V, V_1, V_2 niech $\mathcal{C}_1$ będzie podziałem V_1, niech $\mathcal{C}_2$ będzie podziałem V_2, niech $\mathcal{C}', \mathcal{C}''$ będą podziałami V. Niech będą dane grafy $G_1 = (V \cup V_1, E_1, L, l_1)$ i $G_2 = (V \cup V_2, E_2, L, l_2)$ takie, że E_1, l_1 E_2, l_2 są tożsame nad zbiorem V, to jeśli $Q(G_1, \mathcal{C}' \cup \mathcal{C}_1) \geq Q(G_1, \mathcal{C}'' \cup \mathcal{C}_1)$, to także $Q(G_2, \mathcal{C}' \cup \mathcal{C}_2) \geq Q(G_2, \mathcal{C}'' \cup \mathcal{C}_2)$ (lokalność, uogólnienie bezskalowości rozdzielczości)

Agenda

1. Czym są wspólnoty - wspólnoty i grafy
2. Wymagania na funkcję oceny jakości skupień jako formalne wymaganie wobec algorytmów
3. Jakość skupień w praktyce
   - Rozcięcie grafu, jego koszt
   - Lokalna jakość wspólnoty
   - Globalna jakość wspólnoty
4. Algorytmy
5. Podsumowanie

Rozcięcie grafu

Zakłada się tabelaryzację funkcji I w postaci macierzy podobieństwa S: $I(i,j) = s_{ij}$.
Parę $\{C, \overline{C}\}$, gdzie $C \subset V$ i $\overline{C} = V \backslash C$ nazwiemy rozcięciem grafu.
koszt tego rozcięcia to

$$cut(C, \overline{C}) = R(C, \overline{C}) = \sum_{\substack{v_i \in C \\ v_j \in \overline{C}}} s_{ij} \tag{1}$$

Warianty realnych funkcji oceny jakości skupień w oparciu o koszt rozcięcia

$$Mcut(C_1, \ldots, C_k)) = \sum_{i=1}^{k} cut(C_i, \overline{C}_i) \tag{2}$$

$$Ncut(C_1, \ldots, C_k)) = \sum_{i=1}^{k} \frac{cut(C_i, \overline{C}_i)}{\mathrm{vol} C_i} \tag{3}$$

$$Rcut(C_1, \ldots, C_k)) = \sum_{i=1}^{k} \frac{cut(C_i, \overline{C}_i)}{|C_i|} \tag{4}$$

$$MinMaxcut(C_1, \ldots, C_k) = \sum_{i=1}^{k} \frac{cut(C_i, \overline{C}_i)}{assoc(C_i)} \tag{5}$$

Mcut - (nienormalizowana) funkcja kosztu rozcięcia.
Ncut, ***Rcut*** oraz ***MinMaxcut*** to normalizowana, ilorazowa, oraz min-maxowa funkcja kosztu rozcięcia.

$$assoc(Z) = \sum_{v_i, v_j \in Z} s_{ij} = \mathrm{vol} Z - cut(Z, \overline{Z}) \tag{6}$$

Logiczne warianty lokalne jakości wspólnoty

- silna wspólnota: dla każdego członka wspólnoty suma podobieństw do innych elementów wspólnoty wyższa niż suma podobieństw do nieczłonków
- słaba wspólnota: suma podobieństw wszystkich członków wspólnoty do innych elementów wspólnoty wyższa niż suma ich podobieństw do nieczłonków

Podobieństwo członków wspólnoty jako miara jakości wspólnoty

- suma kwadratów różnic podobieństw parami członków wspólnoty do wszystkich elementów $\boldsymbol{V}$.
- korelacja Pearsona podobieństw parami członków wspólnoty do wszystkich elementów $\boldsymbol{V}$.
- współczynnik Jaccarda (wspólność otoczeń wierzchołków)
- liczba wszystkich ścieżek łączących elementy

Modularność Newmana (odstawanie od losowego przydziału krawędzi) jako miara jakości podziału na wspólnoty

$$Q(S, C) = \frac{1}{2s} \sum_{i=1}^{m} \sum_{j=1}^{m} (s_{ij} - \frac{s_i s_j}{2s}) \delta(C(i), C(j)) \tag{7}$$

gdzie

$$s_i = \sum_{j=1, j \neq i}^{m} s_{ij} \tag{8}$$

and

$$2s = \sum_{i=1}^{m} s_i = \sum_{i=1}^{m} \sum_{j=1, j \neq i}^{m} s_{ij} \tag{9}$$

Agenda

k-medoids

- Oparty o niepodobieństwa
- "centrami" skupień mogą być tylko elementy zbioru
- PAM, CLARA, CLARANS

Ala-jądrowe podejście

- Oparty o niepodobieństwa
- Macierz odległości $\boldsymbol{D}$ winna być dodatnio(pół)określona
- Liczymy jej rozkład na wektory własne $\boldsymbol{V\Lambda V^T}$, gdzie $\boldsymbol{V}$ to tablica wektorów własnych, a $\boldsymbol{\Lambda}$ to macierz diagonalna wartości własnych (które są nieujemne).
- Macierz $\boldsymbol{V\sqrt{\Lambda}}$ traktujemy jako macierz danych i grupujemy metodą $\boldsymbol{k}$-means

Zamiast całej macierzy $\boldsymbol{V\sqrt{\Lambda}}$ można użyć podzbiór wektorów własnych z wartościami własnymi największymi.

Liczą się skupienia, zaś centra skupień są bez merytorycznego znaczenia

k-means oparty na wektorach własnych macierzy podobieństwa

- Oparty o macierz podobieństwa S, i jej podsumowanie D – macierz diagonalna, z elementami będącymi sumami wierszy S.
- przybliżanie normalizowanych przecięć
- liczymy macierz Laplasjanu $L = D - S$ lub normalizowanego Laplasjanu $L' = D^{-}1L$
- Liczymy własności i wektory własne L lub L' i konstruujemy tablicę z k wektorów własnych odpowiadających k najmniejszym wartościom własnym
- Stosujemy k-means na tej macierzy
- Przybliża się w ten sposób nienormalizowane/normalizowane rozcięcie na k podgrafów.

Wektor Fiedlera

Do podziału na dwie grupy używa się wektora odpowiadającego 2giej najmniejszej (pierwszej pozytywnej wartości dla grafów spójnych) wartości własnej. Jedna grupa dla dodatnich druga ujemnych wartości tego wektora

Metody błądzenia losowego

- Oparty o macierz prawdopodobieństwa przejścia (ala-PageRank)
- "Rozpływanie" PageRanku z wybranego miejsca – obszar "rozpłynięcia" skupieniem
- ograniczona liczba kroków rozpływania, z preferencją tzw. "leniwego" PageRanku – Nibble oraz PageRank-Nibble
- inny wariant : MCL - "gwałtowne" chodzenie uniwersalnego globalnego wędrowca (start równomierny)
- jeszcze inny wariant: "męczące się" rozchodzenie się począwszy od kilku inicjalnych centrów i przypisywanie do "zwycięskiego" centrum

Metody optymalizacji modularności – algorytm Louvain

- każdy wierzchołek grafu początkowo własnym skupieniem
- jak długo się da, przesuwamy wierzchołki ze skupienia do skupienia, o ile (maksymalnie) podwyższają modularność.
- Gdy aktualne wierzchołki nie dają podwyższenia modularności, jednostkami przesyłanymi (superwęzłami) stają się aktualne skupienia i tak postępujemy jak poprzednio.

Algorytm jest zgrubny.
Wiele modyfikacji, np. przesuwające tylko węzły, które sąsiadują ze skupieniami

Metody wąskiego gardła informacyjnego

- Przybliżenie koncepcji kosztu rozcięcia
- Wzajemna informacja zmiennej X i Y
  $I(X;Y) = \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$.
- analiza skupień postrzegana jako metoda kompresji danych tak aby reprezentacja skupienia $\hat{X}$ mogła być użyta zamiast pierwotnych danych – czyli choć wzajemna informacja $I(\hat{X},Y) \leq I(X;Y)$, to strata winna być "akceptowalna". Stąd minimalizacja:

$$\mathcal{L}[p(\hat{x}|x)] = I(\hat{X},X) - \beta I(\hat{X},Y) \tag{10}$$

  gdzie β to parametr użytkownika.
- Tu $p(x,y)$ jest dobierane tak: zakładamy rozkład brzegowy $p(x)$, natomiast $p(y|x)$ jest prawdopodobieństwem dotarcia do wierzchołka y startując z x w t krokach w błądzeniu losowym.
- HAK: gdy t zmierza do nieskończoności w grafie spójnym zmierzamy do rozkładu stacjonarnego, i y nie zależy od x.
- WYJŚCIE: wędrowanie z "męczeniem się"? Czy t można zastąpić racjonalnie przez prędkość męczenia?

Agenda

Wyzwania

- Brak jednorodnej koncepcji wspólnoty czy choćby skupienia
- Wielość miar oceny jakości skupień / wspólnot – brak studium, kiedy są one praktycznie równoważne (w tym z powodu braku odwzorowania podobieństwo/odmienność)
- Wielość algorytmów (potęgowana przez ich mieszaniny), niedoskonale optymalizujących te miary – brak studium, co tak naprawdę dostarczają te algorytmy (dla jakiego rodzaju skupień mają one sens)
- Brak studium, kiedy algorytmy te dostarczą równoważnych wyników
- Rozdźwiek między teoretycznymi własnościami algorytmów a realnie stosowanymi algorytmami
- Brak koncepcji i benchmarków dla porównania algorytmów w przypadkach relewantnych praktycznie
- Brak koncepcji próbkowania grafu w zadaniach klasteryzacji

Nieśmiała propozycja

- *Kryterium oceny jakości 1*: Koszt i zyski fizycznej klasteryzacji obiektów
- *Kryterium oceny jakości 2*: Koszt, konieczność i zyski fizycznej reklasteryzacji obiektów w wypadku przewidywalnych zmian własności (połączeń) obiektów
- *Kryterium oceny jakości 3*: wyuczalność ekonomicznej w powyższym sensie klasteryzacji z próbki
- *Idea próbkowania* – poprzez męczącego się losowego wędrowca.

Dziękuję

Dziękuję za uwagę. Uwagi, pytania, komentarze ???