Eksploracja zasobów internetowych

Wykład 6

Grupowanie stron WWW. Funkcje oceniające.

mgr inż. Maciej Kopczyński

Białystok 2015

Wstęp

Rolą algorytmów grupujących jest pogrupowanie dokumentów na bazie ich podobieństwa lub stworzenie modelu opisującego grupę na podstawie reprezentacji obiektów składających się na grupę. Dzięki temu odzwierciedlona jest tematyka grupy dokumentów.

Następnie można wykorzystać funkcje oceny do określenia jakości grup stworzonych przez poszczególne algorytmy.

Jednak do wykonania oceny niezbędne jest posiadanie zbioru dokumentów, który został stworzony (opisany) przez eksperta.

Wstęp

Funkcje oceniające wraz ze zbiorami wzorcowymi (stworzonymi przez eksperta) mogą zostać wykorzystane w dwóch aspektach w przypadku grupowania danych:

- określenia miar błędów popełnianych przez algorytmy grupujące,
- rozszerzenia zbiorów stworzonych przez eksperta o nowe elementy, które są podobne pod względem cech do tych, które zostały sklasyfikowane przez eksperta.

Funkcje oceniające

Jedną z najbardziej popularnych funkcji oceniających jest funkcja najmniejszych kwadratów. Jej istota polega na obliczaniu różnic pomiędzy środkami klastrów (centroidami) a poszczególnymi obiektami należącymi do klastrów.

Funkcja najmniejszych kwadratów opisana jest zależnością:

$$J_e = \sum_{i=1}^{k} \sum_{x \in D_i} \|x - m_i\|$$

gdzie m_i jest centroidem klastra D_i:

$$m_i = \frac{1}{|D_i|} \sum_{x \in d_i} x$$

Funkcje oceniające

Jaki rodzaj miary wykorzystywany jest w klasycznej metodzie najmniejszych kwadratów?

W klasycznej metodzie najmniejszych kwadratów wykorzystywana jest miara euklidesowa.

Jaką najbardziej pożądaną wartość powinna osiągać funkcja oceniająca w klasycznej metodzie najmniejszych kwadratów?

Funkcja oceniająca w klasycznej metodzie najmniejszych kwadratów powinna dążyć do wartości równej 0.

Funkcje oceniające

W przypadku grupowania dokumentów najczęściej wykorzystywana jest miara cosinusowa. Funkcja oceniająca dla metody najmniejszych kwadratów dla miary cosinusowej zdefiniowana jest jako:

$$J_s = \sum_{i=1}^{k} \sum_{d_j \in D_i} \mathrm{sim}(c_i, d_j)$$

gdzie *sim*(c_i, d_j) określa podobieństwo pomiędzy centroidem c_i oraz dokumentem d_j:

$$\mathrm{sim}(c_i, d_j) = \frac{c_i \cdot d_j}{\|c_i\| \, \|d_j\|}$$

Funkcje oceniające

Centroid c_i określający środkowy dokument w ramach zadanego klastra zdefiniowany jest jako:

$$c_i = \frac{1}{D_i} \sum_{d_j \in D_i} d_j$$

Jaka jest najbardziej pożądana wartość funkcji oceniającej wykorzystującej miarę cosinusową?

Najbardziej pożądaną wartością funkcji oceniającej wykorzystującej miarę cosinusową jest wartość jak największa.

Funkcje oceniające

Grupowanie, które maksymalizuje wartość funkcji oceniającej jest określane mianem grupowania o minimalnej wariancji.

Funkcje oceniające należące do tej rodziny są określane jako funkcje oceniające o minimalnej wariancji.

Funkcje oceniające mogą również przyjmować dwie formy zależnie od rodzaju grupowania:

- funkcje bazujące na centroidzie,
- funkcje bazujące na odległościach pomiędzy parami obiektów.

Porównanie klastrów

Istotną kwestią jest możliwość porównania między sobą grupowań dokumentów o różnych licznościach klastrów lub różnych strukturach hierarchicznych.

W przypadku grupowania płaskiego, wartość funkcji oceniającej rośnie wraz ze zwiększaniem ilości klastrów.

W przypadku grupowania hierarchicznego wartość funkcji oceniającej rośnie wraz z przemieszczaniem się w dół dendrogramu.

Porównanie klastrów – przykład

Na kolejnym slajdzie pokazane będzie uruchomienie algorytmu grupowania hierarchicznego (*agglomerative hierarchical clustering*) oraz trzy uruchomienia algorytmu *k-średnich* z parametrami $k = 2$, 3 oraz 4.

Wartość funkcji oceniającej na bazie centroidów zapisana jest w nawiasie kwadratowym obok stworzonego klastra. Dany klaster składa się z dokumentów leżących poniżej w hierarchii tworzącej grupę.

Agglomerative	k-means $(k = 2)$	k-means $(k = 3)$	k-means $(k = 4)$
1 [12.0253]	1 [12.0253]	1 [12.0253]	1 [12.0253]
2 [9.43932]	2 [8.53381]	2 [2.83806]	2 [3.81771]
3 [5.64819]	Anthropology	History	Art
4 [4.6522]	Biology	Music	Communication
5 [3.8742]	Chemistry	Philosophy	English
6 [2.95322]	Computer Science	3 [6.09107]	Modern Languages
7 [1.99773]	Economics	Anthropology	3 [5.44416]
Chemistry	Geography	Biology	Biology
Computer Science	Mathematics	Chemistry	Economics
Political Science	Physics	Computer Science	Mathematics
Geography	Political Science	Geography	Physics
Anthropology	Psychology	Mathematics	Psychology
8 [1.98347]	Sociology	Political Science	Sociology
Criminal Justice	3 [6.12743]	4 [7.12119]	4 [2.83806]
Theatre	Art	Art	History
9 [5.44416]	Communication	Communication	Music
10 [2.81679]	Criminal Justice	Criminal Justice	Philosophy
11 [1.97333]	English	Economics	5 [5.64819]
Psychology	History	English	Anthropology
Sociology	Modern Languages	Modern Languages	Chemistry
Mathemetics	Music	Physics	Computer Science
12 [2.90383]	Philosophy	Psychology	Criminal Justice
13 [1.96187]	Theatre	Sociology	Geography
Biology		Theatre	Political Science
Economics			Theatre
Physics			
14 [5.40061]			
15 [2.83806]			
16 [1.98066]			
History			
Music			
Philosophy			
17 [3.81771]			
18 [2.97634]			
19 [1.99175]			
English			
Modern Languages			
Art			
Communication			
14.83993 (clusters 2 + 14)	**14.6612**	**16.05032**	**17.74812**

Porównanie klastrów – przykład

W algorytmie grupowania hierarchicznego, węzeł nr 3 reprezentuje klaster składający się z następujących dokumentów: *Chemistry*, *Computer Science*, *Political Science*, *Geography*, *Anthropology*, *Criminal Justice* oraz *Theatre*. Funkcja oceniająca bazująca na centroidach zwróciła wynik 5,64819.

Suma wartości w węzłach pochodnych jest różna od wartości 5,64819.

Dzieje się tak dlatego, że funkcja ocenia pojedyncze klastry bez uwzględniania zależności w drzewie.

Porównanie klastrów – przykład

Dzięki temu można porównać ze sobą dwa klastry znajdujące się na różnych poziomach drzewa w przypadku algorytmu hierarchicznego lub dwa klastry o różnych licznościach w przypadku algorytmu grupowania płaskiego.

Za pomocą funkcji oceniających można przeanalizować problem łączenia ze sobą poszczególnych klastrów. Przy analizie algorytmu hierarchicznego, najpierw należy zidentyfikować klastry, które na najniższym poziomie pokrywają cały zbiór dokumentów. Są to klastry: 4, 8, 10, 12, 15 oraz 17. Teraz, korzystając z kombinacji łączeń tych klastrów, można wybrać ten sposób połączenia klastrów, który da największą wartość funkcji oceniającej.

Łączenie klastrów

Dla klastrów o węzłach podanych na poprzednim slajdzie, można wygenerować następujące kombinacje łączeń:

Partitioning	Sum of Centroid Similarity
{2,14}	14.8399
{3,9,14}	16.493
{2,15,17}	16.0951
{4,8,9,14}	17.4804
{3,9,15,17}	17.7481
{4,8,9,15,17}	18.7356
{3,10,12,15,17}	18.0246
{4,8,10,12,14}	17.7569
{4,8,10,12,15,17}	19.0121

Z tabeli kombinacji wynika, że im więcej klastrów powstaje w wyniku grupowania, tym większa wartość funkcji oceniającej.

Łączenie klastrów

Intuicja podpowiada jednak, że chcemy tworzyć jak największe klastry. Jednak przy łączeniu klastrów ze sobą spada wartość funkcji oceniającej.

Tak więc istotny jest kompromis pomiędzy wartością funkcji oceniającej, a wielkością klastrów.

Można to osiągnąć poprzez:

- znajomość tematyki związanej z danymi klastrami (np. łączenie 4, 8, 9 oraz 14),
- przy braku znajomości tematyki warto łączyć klastry na tych samych poziomach (np. 3, 9 oraz 14).

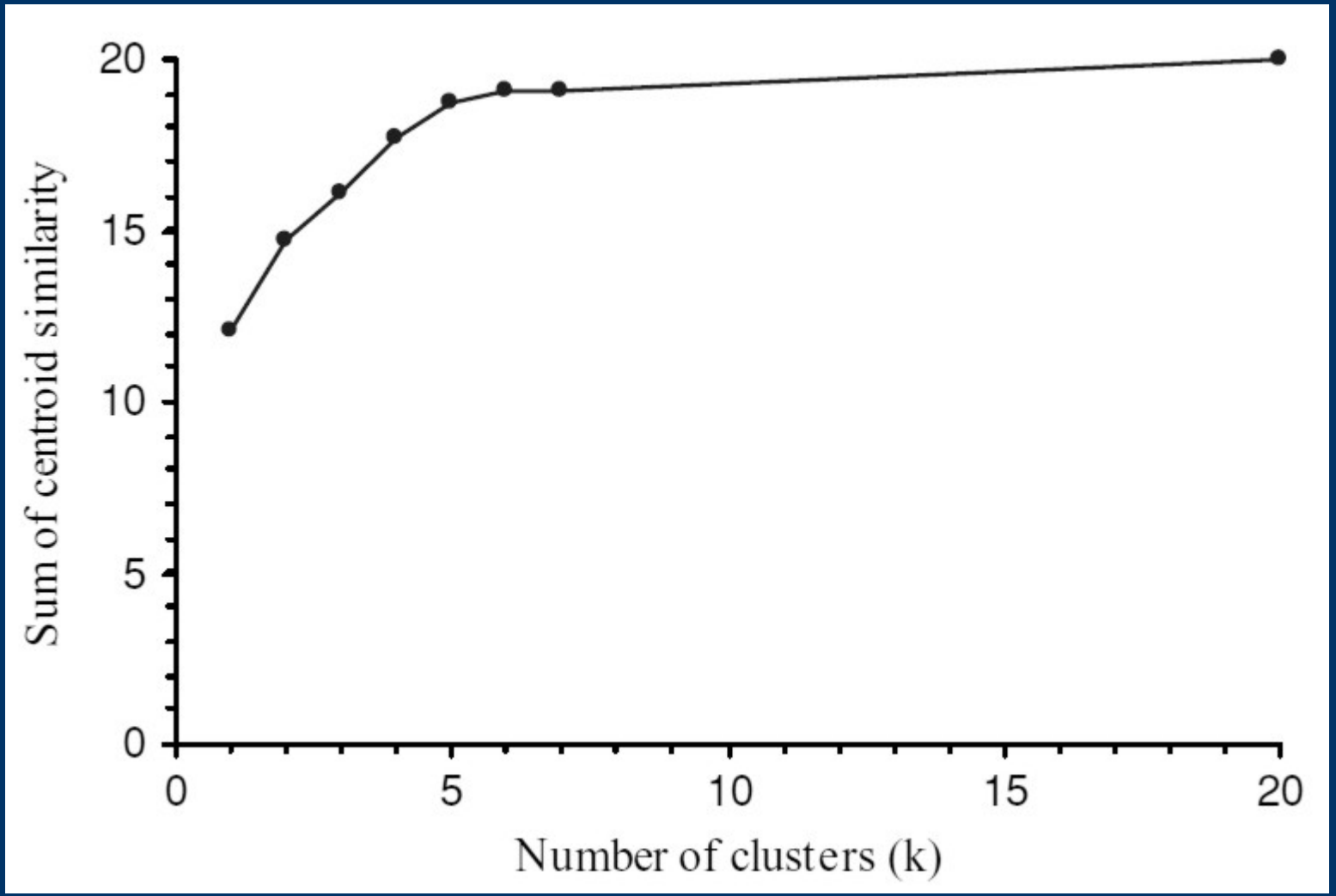
Łączenie klastrów

W przypadku algorytmu *k-średnich* można zastosować inną technikę wyszukiwania optymalnej ilości klastrów.

Najpierw wykonane zostały trzy kolejne uruchomienia algorytmu *k-średnich* dla parametru k = 5, 6, 7 oraz 20 (ilość dokumentów w zbiorze). Dla każdego uruchomienia obliczona została wartość funkcji oceniającej bazującej na centroidach.

Na kolejnym slajdzie pokazany jest wykres wartości funkcji oceniającej w zależności od parametru k (ilości klastrów).

Łączenie klastrów



Łączenie klastrów

Dla pierwszych wartości parametru *k* wykres jest stromy, zaś powyżej tej wartości wykres staje się płaski.

Głównym celem algorytmu *k-średnich* jest maksymalizacja wartości funkcji oceniającej.

Czym będzie skutkowała nadmierna maksymalizacja wartości funkcji oceniającej?

Nadmierna maksymalizacja wartości funkcji będzie skutkowała zbyt dużą ilością klastrów wynikowych.

Łączenie klastrów

Który w takim razie punkt na wykresie jest najbardziej optymalny pod względem ilości klastrów i wartości funkcji oceniającej?

Najbardziej optymalnym punktem na wykresie pod względem ilości klastrów i wartości funkcji oceniającej jest miejsce, w którym wykres przechodzi z szybkiego wzrostu do części płaskiej.

W przypadku danych przykładowych, taki punkt występuje dla $k = 5$. Rozwiązanie to jest poprawne w przypadku, gdy nie mamy żadnych informacji na temat zbioru.

W przypadku naszego zbioru danych najbardziej pożądana jest war-
parametru $k = 2$.

Wyszukiwanie wzorców

Do tej pory grupowanie omawiane było jako metoda tworzenia klastrów składających się z dokumentów o podobnych cechach.

Grupowanie wraz z funkcjami oceniającymi może być również wykorzystane jako mechanizm do odnajdowania wzorców. Jeśli dany klaster składa się z względnie dużej ilości obiektów oraz wartość jego funkcji oceniającej również jest wysoka, może to świadczyć o odnalezieniu klastra reprezentującego jakiś wzorzec.

Na podstawie klastra można stworzyć reguły opisujące wzorzec.

Wyszukiwanie wzorców

Wzorzec będzie opisany przez zbiór reguł, w których warunki są zdefiniowane przez atrybuty posiadające takie same wartości dla wszystkich dokumentów należących do danego klastra.

Zbiór przykładowych reguł może wyglądać następująco:

- R_1: IF (*science* = 0) AND (*research* = 1) THEN class = A
- R_2: IF (*science* = 1) AND (*research* = 0) THEN class = A
- R_3: IF (*science* = 1) AND (*research* = 1) THEN class = A
- R_4: IF (*science* = 0) AND (*research* = 0) THEN class = B

Wyszukiwanie wzorców

Reguły mogą być później wykorzystane do klasyfikacji (grupowania) innych zbiorów danych.

Przy wykorzystaniu reguły *dropping conditions* można uogólniać wzorzec poprzez usuwanie elementów części warunkowej reguł.

Tworzenie nadmiernie dokładnych reguł, a także zbytnio ogólnych reguł jest niepożądane ze względu na generowanie dwóch skrajnych rodzajów grup.

Przy zbyt ogólnych regułach, wszystkie elementy trafią do jednego klastra. Przy nadmiernie szczegółowych regułach obiekty stworzą jednoelementowe klastry.

Miary oceny

Załóżmy, że dla każdego z dokumentów na bazie jego zawartości została przyporządkowana jedna z dwóch klas: *A* lub *B*.

Następnie dokumenty zostały pogrupowane ze względu na występowanie danego termu w ich treści. Tego typu grupowanie pozwala na pomiar błędów jakości grupowania oraz błędów doboru odpowiednich atrybutów.

Kolejny slajd przedstawia grupowanie ze względu na kilka wybranych termów. Dla każdej z grup podane są wartości błędów.

history 14/20		*science* 16/20		*research* 17/20		*offers* 14/20		*students* 14/20		*hall* 12/20	
A (11/17)	*B* (3/3)	*B* (9/13)	*A* (7/7)	*B* (9/12)	*A* (8/8)	*A* (8/11)	*B* (6/9)	*B* (4/5)	*A* (10/15)	*B* (3/5)	*A* (9/15)
1-A	11-B	2-B	1-A	2-B	1-A	1-A	2-B	8-A	1-A	5-B	1-A
2-B	14-B	5-B	3-A	4-A	3-A	3-A	5-B	9-B	2-B	7-B	2-B
3-A	15-B	7-B	4-A	5-B	8-A	4-A	9-B	11-B	3-A	8-A	3-A
4-A		8-A	6-A	6-A	12-A	6-A	11-B	13-B	4-A	10-A	4-A
5-B		9-B	10-A	7-B	16-A	7-B	12-A	15-B	5-B	11-B	6-A
6-A		11-B	12-A	9-B	17-A	8-A	13-B		6-A		9-B
7-B		13-B	17-A	10-A	18-A	10-A	15-B		7-B		12-A
8-A		14-B		11-B	19-A	14-B	18-A		10-A		13-B
9-B		15-B		13-B		16-A	19-A		12-A		14-B
10-A		16-A		14-B		17-A			14-B		15-B
12-A		18-A		15-B		20-B			16-A		16-A
13-B		19-A		20-B					17-A		17-A
16-A		20-B							18-A		18-A
17-A									19-A		19-A
18-A									20-B		20-B
19-A											
20-B											

Miary oceny

Miara ilości popełnianych błędów w stosunku do wszystkich dokumentów jest czasami niewystarczająca. W wielu przypadkach warto jest znać rodzaj popełnianych błędów.

Biorąc jako przykład system filtracji antyspamowej wiadomości email, jaki rodzaj błędów popełnianych podczas klasyfikacji jest bardziej istotny?

W systemie antyspamowym można obliczyć ilość wszystkich popełnianych przez system pomyłek, jednak dużo bardziej istotna jest błędna klasyfikacja dobrego emaila jako spam, niż spamu jako dobry email.

Miary oceny

Ze względu na to, że większość problemów grupowania i klasyfikacji operuje na problemie dwóch klas, wprowadzone zostały odpowiednie miary w zależności od typu klasyfikacji dokumentu:

- *True positive* (*TP*) – dokument prawdziwy sklasyfikowany jako prawdziwy,
- *False positive* (*FP*) – dokument nieprawdziwy sklasyfikowany jako prawdziwy,
- *True negative* (*TN*) – dokument nieprawdziwy sklasyfikowany jako nieprawdziwy,
- *False negative* (*FN*) – dokument prawdziwy sklasyfikowany jako nieprawdziwy.

Miary oceny

Najprościej jest przedstawić wprowadzone błędy jako strukturę macierzową związaną z rodzajem klasyfikacji dokumentu:

	Predicted (Clusters)	
Actual (Classes)	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Jaki jest najlepszy przypadek ułożenia danych w tej macierzy?

W najlepszym przypadku wartości niezerowe będą umieszczone tylko na przekątnej macierzy.

Miary oceny

Przykład rozłożenia wyników dla klas *A* i *B* oraz termów *research* oraz *hall*:

	research		*hall*	
	Predicted (Clusters)		Predicted (Clusters)	
Actual (Classes)	*A*	*B*	*A*	*B*
A	8	3	9	2
B	0	9	6	3

Term *research* posiada lepiej rozłożone współczynniki niż term *hall*.

Miary oceny

Błąd ogólny z wykorzystaniem zdefiniowanych współczynników jest zdefiniowany jako:

$$error = \frac{FP + FN}{TP + FP + TN + FN}$$

Ogólna dokładność klasyfikacji jest zdefiniowana jako:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Miary oceny

Podobnie jak w przypadku techniki IR, tak dla grupowania można zdefiniować współczynniki *precision* oraz *recall*.

Współczynnik *precision* jest zdefiniowany jako:

$$precision = \frac{TP}{TP + FP}$$

Wartości współczynnika *precision* zawierają się w zakresie od 0 do 1.

Współczynnik opisuje stosunek obiektów należących do klasy *A* do obiektów należących do klasy *A* w rzeczywistości oraz wg algorytmu.

Miary oceny

Współczynnik *recall* jest zdefiniowany jako:

$$recall = \frac{TP}{TP + FN}$$

Wartości współczynnika *recall* zawierają się w zakresie od 0 do 1.

Współczynnik opisuje stosunek obiektów należących do klasy *A* do obiektów należących do klasy *A* w rzeczywistości oraz należących do klasy *B* wg algorytmu.

Miary oceny

Współczynnik *recall* może być sprowadzony do wartości równej 1 dla przypadku, gdy wszystkie dokumenty zostaną sklasyfikowane jako obiekty należące do klasy *A*.

Współczynnik *precision* może być sprowadzony do wartości równej 1 dla przypadku, gdy klasyfikacja dotyczyła będzie tylko jednego dokumentu oraz ten dokument zostanie sklasyfikowany jako dokument należący do klasy *A*.

W idealnym przypadku współczynniki powinny być jednocześnie równe 1.

Miary oceny

Dla termu *research* współczynnik *precision* jest równy 1,0, zaś *recall* ma wartość 0,73. Dla termu *hall* współczynnik *precision* jest równy 0,6, zaś *recall* ma wartość 0,82.

Zależnie od rozpatrywanego problemu, jeden ze współczynników można uznać za priorytetowy. W przypadku systemu antyspamowego, gdzie klasa *positive* będzie oznaczała istotną wiadomość email, to drugi rodzaj klasyfikacji jest lepszy, gdyż przepuszcza więcej wiadomości istotnych (82%), pomimo tego, że współczynnik *precision* dla pierwszego rodzaju klasyfikacji przyjmuje wartość 100%.

Podsumowanie

Funkcje oceniające są nieodłącznym elementem grupowania danych, a w szczególności dokumentów sieci WWW. Pozwalają oceniać jakość grupowania oraz mogą być wykorzystane do optymalizacji wartości parametrów określających ilości grup tworzone przez algorytmy.

Funkcje oceniające mogą przyjmować różne postacie: bazujące na centroidach, bazujące na podobieństwie par obiektów czy też korzystające z przyporządkowania obiektów do klas.

Dziękuję za uwagę!