PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Jerzy Korzeniewski
Uniwersytet Łódzki

BADANIE EFEKTYWNOŚCI ZMODYFIKOWANYCH METOD AGLOMERACYJNYCH NA ZBIORACH DANYCH ZE ŚWIATA REALNEGO

1. Wstęp

Niniejszy artykuł zawiera wyniki badania efektywności zmodyfikowanych metod aglomeracyjnych służących do grupowania obserwacji. Ideą modyfikacji algorytmów aglomeracyjnych jest wprowadzenie do stosowanego kryterium dodatkowego pomiaru mającego na celu łączenie skupień z uwzględnieniem lokalnej gęstości rozkładu obserwacji, tj. zwiększenie znaczenia obszarów o dużej gęstości obserwacji przy ustalaniu kolejności łączenia istniejących skupień. W dalszym ciągu artykułu znajduje się dokładne sformułowanie modyfikacji kilku wybranych algorytmów aglomeracyjnych, omówienie metody ustalania liczby skupień, na jaką należy podzielić zbiór danych, i metody oceny jakości podziału oraz wyniki badania kilku zbiorów danych ze świata realnego.

2. Sformułowanie zmodyfikowanych metod

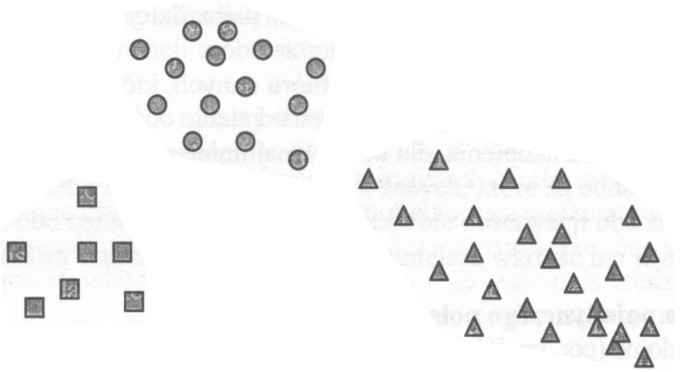
Jak wiadomo, hierarchiczne metody aglomeracyjne charakteryzują się następującymi cechami (por. np. [Gordon 1999]):

- punktem wyjścia jest n klas jednoelementowych (tzn. tyle, ile obserwacji w zbiorze),
- w każdym kroku aglomeracji liczba klas zmniejsza się o jeden, przy czym zmniejszenie liczby klas następuje przez połączenie dwóch klas istniejących,
- po $n-1$ krokach otrzymujemy jedną klasę zawierającą wszystkie obserwacje.

Łączenie klas, czyli grupowanie obserwacji, odbywa się przez zastosowanie następującego algorytmu.

- W macierzy odległości między klasami szuka się pary klas najbardziej podobnych w sensie przyjętego kryterium (np. najmniej odległych). Załóżmy, że będą to klasy o numerach *i, j*.
- Zmniejsza się liczbę klas o jeden, łącząc klasy o numerach *i, j*.
- Przekształca się macierz odległości między klasami, tak by znów były zdefiniowane wszystkie pary odległości (definiuje się odległości między nowo powstałą klasą a pozostałymi klasami).
- Powyższe trzy kroki powtarza się do momentu, aż wszystkie obiekty znajdą się w jednej klasie.

Podstawową wadą takiego algorytmu jest wada „łańcucha". Polega ona na tym, że algorytm ten, posługując się różnie definiowanymi odległościami między skupieniami, ma tendencję do łączenia skupień najbliższych, w efekcie czego do tej samej klasy można zaliczyć bardzo różne obiekty zbioru, ale „połączone" łańcuchem obiektów, z których każde dwa kolejne są mało zróżnicowane, tj. bliskie w sensie przyjętej odległości. Można spróbować wyeliminować tę wadę, preferując w każdym kroku algorytmu łączenie skupień znajdujących się w rejonie, w którym gęstość rozkładu obiektów jest większa. Prześledźmy ten pomysł na ilustracji (zob. rys. 1).



Rys. 1. Trzy skupienia obserwacji z dwuwymiarowej przestrzeni euklidesowej (każde zaznaczone innym kształtem)

Źródło: opracowanie własne.

Załóżmy, że mamy dokonać połączenia dwóch spośród trzech skupień widocznych na rys. 1. Gdybyśmy chcieli zastosować algorytm aglomeracyjny łączący dwa skupienia o najbliższych środkach ciężkości, to wypadałoby połączyć skupienia reprezentowane przez koła i kwadraty. Jeśli wprowadzimy modyfikację polegającą na uwzględnieniu większej gęstości obserwacji między skupieniami kół i trójkątów, to może się okazać, że korzystniej byłoby połączyć właśnie te dwa skupienia. Gęstość lokalną rozkładu obserwacji można uwzględnić różnie, naturalne wydaje się podejście polegające na odniesieniu liczby obserwacji znajdujących się w określonym podzbiorze przestrzeni euklidesowej do objętości tego podzbioru. W kontekście przykładu widocznego na rys. 1 i metody

środków ciężkości mogłoby to wyglądać następująco: łączymy te dwa skupienia, dla których najmniejsza jest odległość ich środków ciężkości podzielona przez liczbę punktów znajdujących się w odległości mniejszej od odległości środków ciężkości od każdego ze środków. Dodatkowo należałoby odnieść tę liczbę punktów (tj. podzielić ją przez) do objętości odnośnego podzbioru płaszczyzny, czyli w omawianym przypadku do wyrażenia proporcjonalnego do kwadratu odległości środków ciężkości. Kwestią do wyboru pozostaje wyznaczanie podzbioru przestrzeni euklidesowej, z którego będziemy zliczali punkty, w przypadku bowiem niektórych algorytmów aglomeracyjnych (np. metody średniej odległości klasowej) nie ma naturalnie wyznaczonych choćby dwóch punktów odniesienia, takich jak środki ciężkości w przykładzie z rys. 1. Wówczas należy zaproponować takie punkty odniesienia. Poniżej znajdują się dokładne sformułowania czterech zmodyfikowanych metod aglomeracyjnych.

Metoda całkowitego połączenia

Jak wiadomo (por. np. [Gordon 1999]), metoda ta działa w ten sposób, że w każdym kroku algorytmu łączymy te dwa skupienia, które mają najmniejszą odległość dwóch najbardziej oddalonych od siebie obserwacji tych skupień. Modyfikację definiujemy następująco.

- Znajdujemy odległość r między dwoma najbardziej odległymi obserwacjami dla każdej pary skupień.
- Znajdujemy liczbę x obserwacji ze zbioru danych, które są oddalone o nie więcej niż r od obu najbardziej oddalonych od siebie obserwacji obu skupień.
- Łączymy te dwa skupienia, dla których najmniejszą wartość ma wyrażenie

$$\frac{r}{x/r^d} = \frac{r^{d+1}}{x}.$$

Metoda pojedynczego połączenia

Jak wiadomo (por. np. [Gordon 1999]), metoda ta działa w ten sposób, że w każdym kroku algorytmu łączymy te dwa skupienia, które mają najmniejszą odległość dwóch najbliższych sobie obserwacji tych skupień. Modyfikację definiujemy następująco.

- Znajdujemy odległość r między dwoma najbardziej odległymi obserwacjami dla każdej pary skupień.
- Znajdujemy liczbę x obserwacji ze zbioru danych, które są oddalone o nie więcej niż r od obu najbardziej oddalonych od siebie obserwacji obu skupień.
- Łączymy te dwa skupienia, dla których najmniejszą wartość ma wyrażenie

$$\frac{s}{x/r^d} = \frac{s \cdot r^d}{x},$$

gdzie: s – odległość między dwoma najbliższymi obserwacjami obu skupień.

W powyższej modyfikacji nie ma zliczania obserwacji między dwoma najbliższymi obserwacjami dwóch skupień. Taka modyfikacja niewiele wniosłaby do

znanej metody, bo obserwacji spełniających ten warunek jest mało. Zamiast tego zaproponowano, by x oznaczało liczbę obserwacji leżących „pomiędzy" najbardziej oddalonymi obserwacjami skupień.

Metoda środka ciężkości

Ta metoda działa w ten sposób (por. np. [Gordon 1999]), że w każdym kroku algorytmu łączymy te dwa skupienia, które mają najmniejszą odległość między środkami ciężkości obu skupień. Modyfikację definiujemy następująco.

- Znajdujemy odległość r między dwoma środkami ciężkości dla każdej pary skupień.
- Znajdujemy liczbę x obserwacji ze zbioru danych, które są oddalone o nie więcej niż r od obu środków ciężkości.
- Łączymy te dwa skupienia, dla których najmniejszą wartość ma wyrażenie

$$\frac{r}{x/r^d} = \frac{r^{d+1}}{x}.$$

Metoda średniej klasowej

Metoda ta działa tak (por. np. [Gordon 1999]), że w każdym kroku algorytmu łączymy te dwa skupienia, które mają najmniejszą średnią arytmetyczną odległości dla wszystkich par obserwacji z obu skupień. Modyfikację definiujemy następująco.

- Znajdujemy odległość r między dwoma najbardziej odległymi obserwacjami dla każdej pary skupień.
- Znajdujemy liczbę x obserwacji ze zbioru danych, które są oddalone o nie więcej niż r od obu najbardziej oddalonych od siebie obserwacji obu skupień.
- Łączymy te dwa skupienia, dla których najmniejszą wartość ma wyrażenie

$$\frac{s}{x/r^d} = \frac{s \cdot r^d}{x},$$

gdzie: s – średnia arytmetyczna odległości między wszystkimi parami obserwacji obu skupień.

W powyższej modyfikacji zaproponowano, by x oznaczało liczbę obserwacji leżących „pomiędzy" najbardziej oddalonymi obserwacjami skupień. Można oczywiście zaproponować inną regułę uwzględniania gęstości lokalnej, gdyż w przypadku metody średniej klasowej nie ma konkretnego odniesienia do podzbiorów przestrzeni, z której pochodzą obserwacje.

3. Eksperyment badawczy

Założenia eksperymentu

W celu oceny zaproponowanych modyfikacji wybrano do badania kilka zbiorów danych analizowanych w pracy [Ripley 1996]. Te zbiory są dostępne na stronach

Międzynarodowej Federacji Towarzystw Klasyfikacyjnych (International Federation of Classification Societies): http://lib.stat.cmu.edu/. Zbiory te zostały pogrupowane każdą z ośmiu (cztery znane i ich modyfikacje) metod aglomeracyjnych w rozłączne skupienia w liczbie od 2 od 12. Następnie, w celu ustalenia właściwej liczby skupień, dla każdego zbioru wyznaczono liczbę skupień sugerowaną przez cztery popularne indeksy skupień: Calińskiego-Harabasza, Krzanowskiego-Lai, Hartigana oraz sylwetkowy. Formuły tych indeksów można znaleźć np. w pracy [Sugar, James 2003].

Po ustaleniu najlepszej liczby skupień, sugerowanej przez każdy z czterech indeksów skupień, dla każdej z ośmiu badanych metod aglomeracyjnych ustalana była ostateczna liczba skupień, na które dzielony był zbiór danych. Podstawowym kryterium ustalania liczby skupień było najczęstsze powtarzanie się jakiejś liczby, zwracano także uwagę na to, by ostateczna liczba skupień była bliska średniej arytmetycznej wszystkich pokazywanych liczb. Ostateczna liczba skupień była taka sama dla wszystkich ośmiu metod aglomeracyjnych, niezależnie od tego, że dla niektórych metod być może lepsza byłaby inna liczba skupień.

Po podzieleniu zbioru danych na wcześniej ustaloną liczbę skupień oceniana była jakość podziału za pomocą odsetka źle przypisanych do skupień obserwacji, tj. obserwacji z ujemnym indeksem sylwetkowym (por. np. [Sugar, James 2003]) oraz średniej arytmetycznej wartości wszystkich indeksów ujemnych. Najlepszym zatem podziałem będzie ten podział, który da najmniejszy odsetek źle przypisanych obserwacji i o jak największej (ujemnej), czyli najbliższej zeru średniej wartości tych indeksów.

Charakterystyka zbiorów danych

Wszystkie badane zbiory danych zostały znormalizowane oddzielnie względem każdej zmiennej przez odjęcie od wartości j-tej zmiennej dla i-tej obserwacji średniej arytmetycznej j-tej zmiennej i podzielenie przez odchylenie standardowe tej zmiennej.

Zbiór pierwszy to 214 próbek szkła scharakteryzowanych przez 9 zmiennych. Analiza ustalająca najlepszą liczbę skupień dla tego zbioru dała wyniki przedstawione w tab. 1.

Tabela 1. Liczby skupień wskazywane przez cztery indeksy skupień dla pierwszego i drugiego zbioru danych

Metoda		Zbiór 1				Zbiór 2			
		Syl-wetk.	Cal.--Har.	Hart.	Krzan.--Lai	Syl-wetk.	Cal.--Har.	Hart.	Krzan.--Lai
Środka ciężkości	zmodyfikowana	3	4	5	3	2	5	5	5
	klasyczna	2	5	5	5	2	6	-	6
Środka ciężkości	zmodyfikowana	2	3	3	5	2	5	3	5
	klasyczna	2	4	2	4	2	2	2	3
Średniego połączenia	zmodyfikowana	6	5	3	5	4	4	4	4
	klasyczna	2	3	3	5	2	5	5	5
Pojedynczego połączenia	zmodyfikowana	5	6	6	6	2	6	4	6
	klasyczna	2	2	3	5	3	4	2	6

Źródło: obliczenia własne.

Na podstawie liczb zawartych w tab. 1 liczbę skupień ustalono jako równą 5. Co prawda, formalnie rzecz biorąc, w zbiorze tym było 7 rodzajów szkła, jednak takiej liczby skupień żaden indeks nie pokazał. Wobec tego wybrano najbliższą 7 liczbę często wskazywaną, tj. 5.

Zbiór drugi to 200 okazów dwóch gatunków krabów scharakteryzowanych przez 5 zmiennych. Analiza ustalająca najlepszą liczbę skupień dla tego zbioru dała wyniki przedstawione w tab. 1. Na podstawie tych danych i zgodnie z przyjętą zasadą ustalania liczby skupień, przyjęto liczbę skupień równą 4.

Tabela 2. Liczby skupień wskazywane przez cztery indeksy skupień dla trzeciego i czwartego zbioru danych

Metoda		Zbiór 3				Zbiór 4			
		Syl-wetk.	Cal.--Har.	Hart.	Krzan.--Lai	Syl-wetk.	Cal.--Har.	Hart.	Krzan.--Lai
Środka ciężkości	zmodyfikowana	2	4	6	4	2	5	2	3
	klasyczna	2	5	-	5	2	2	5	3
Środka ciężkości	zmodyfikowana	2	5	2	5	3	3	4	3
	klasyczna	2	2	2	4	2	4	2	6
Średniego połączenia	zmodyfikowana	2	2	3	6	2	5	5	6
	klasyczna	2	3	3	6	3	3	3	4
Pojedynczego połączenia	zmodyfikowana	2	5	2	4	3	2	4	3
	klasyczna	2	3	4	4	3	2	2	5

Źródło: obliczenia własne.

Zbiór trzeci to 250 sztucznie wygenerowanych obiektów scharakteryzowanych przez 2 zmienne. Analiza ustalająca najlepszą liczbę skupień dla tego zbioru dała wyniki przedstawione w tab. 2. Na podstawie tych liczb przyjęto ostateczną liczbę skupień równą 2.

Tabela 3. Liczby skupień wskazywane przez cztery indeksy skupień dla piątego i szóstego zbioru danych

Metoda		Zbiór 5				Zbiór 6			
		Syl-wetk.	Cal.--Har.	Hart.	Krzan.--Lai	Syl-wetk.	Cal.--Har.	Hart.	Krzan.--Lai
Środka ciężkości	zmodyfikowana	2	3	2	4	2	3	3	4
	klasyczna	2	3	3	2	2	3	3	3
Środka ciężkości	zmodyfikowana	2	5	2	5	2	2	2	5
	klasyczna	2	2	2	6	2	6	3	5
Średniego połączenia	zmodyfikowana	3	2	4	3	2	3	3	4
	klasyczna	3	2	2	4	2	2	2	3
Pojedynczego połączenia	zmodyfikowana	3	2	2	3	2	3	3	2
	klasyczna	2	2	3	4	2	3	2	3

Źródło: obliczenia własne.

Zbiór czwarty to 200 Indian ze szczepu Pima scharakteryzowanych przez 7 zmiennych. Analiza liczb zawartych w tab. 2 pozwoliła ustalić 2 jako ostateczną liczbę skupień.

Zbiór piąty to 96 próbek szkła scharakteryzowanych przez 5 zmiennych. Analiza ustalająca najlepszą liczbę skupień dla tego zbioru dała wyniki przedstawione w tab. 3. Na ich podstawie przyjęto ostateczną liczbę skupień równą 2.

Tabela 4. Odsetki obserwacji o ujemnych indeksach sylwetkowych (w kolumnach niezacieniowanych) i średnie arytmetyczne wartości indeksów ujemnych (w kolumnach zacieniowanych pominięto znak minus)

Metoda		Zbiór 1		Zbiór 2		Zbiór 3		Zbiór 4		Zbiór 5		Zbiór 6	
Środka ciężkości	zmodyfikowana	12	0,08	14	0,26	14	0,29	8	0,18	2	0,27	12	0,08
	klasyczna	6	0,17	10	0,22	8	0,30	80,5	0,14	4	0,11	3	0,10
Środka ciężkości	zmodyfikowana	4	0,18	20	0,27	14	0,20	30,5	0,11	3	0,26	6	0,18
	klasyczna	44	0,43	36	0,56	17	0,31	18	0,41	21	0,28	22	0,24
Średniego połączenia	zmodyfikowana	4	0,11	7	0,14	3	0,06	15	0,23	2	0,27	6	0,18
	klasyczna	5	0,17	11	0,23	7	0,32	30,5	0,11	4	0,11	4	0,11
Pojedynczego połączenia	zmodyfikowana	5	0,21	10	0,22	26	0,32	10	0,14	13	0,26	14	0,17
	klasyczna	8	0,17	31	0,45	33	0,34	22	0,23	31	0,30	25	0,23

Źródło: obliczenia własne.

Zbiór szósty to 96 obserwacji dotyczących gospodarki szwajcarskiej scharakteryzowanych przez 4 zmienne. Analiza ustalająca najlepszą liczbę skupień dla tego zbioru dała wyniki przedstawione w tab. 3, z której wynika, że za najlepszą liczbę skupień należałoby uznać liczbę 2.

4. Wnioski końcowe

Liczby zawarte w tab. 4 pozwalają wyciągnąć następujące wnioski:

1. Zaproponowana modyfikacja prawie zawsze poprawia tradycyjny algorytm aglomeracyjny z wyjątkiem algorytmu całkowitego połączenia, który trudno poprawić.
2. Na sześć badanych zbiorów 4 razy najlepsza okazała się zmodyfikowana metoda średniego połączenia, raz klasyczna metoda całkowitego połączenia i raz zmodyfikowana metoda środka ciężkości *ex aequo* z klasyczną metodą średniego połączenia.
3. Zaproponowana modyfikacja nie nadaje się do dużych zbiorów danych – zbyt długi jest czas przeszukiwania tych zbiorów i zliczania obserwacji spełniających określone warunki.

Literatura

Gordon A.D. (1999), *Classification*, Chapman & Hall, London, New York, Washington.

Ripley B.D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.

Sugar C.A., James G.M. (2003), *Finding the Number of Clusters in a Dataset, An Information-theoretic Approach*, „JASA", vol. 98.

INVESTIGATING THE EFFICIENCY OF MODIFIED AGGLOMERATIVE METHODS ON REAL WORLD DATA SETS

Summary

In the paper, the efficiency of agglomerative clustering algorithms is investigated with a special focus on the author's own modifications. The following clustering algorithms are examined: single link, complete link, group average link and centroid link. The idea of the modifications is to stress the local distribution of observations together with clustering based on the dissimilarity matrix. The quality of clustering is assessed by means of the silhouette indices. The results prove that the author's modifications almost always improve the standard methods.