Marcin Wardyński

Analiza skupień dla cars.sta

Metoda k-średnich

Wykres średnich zmiennych ilościowych pokazuje uśrednione, znormalizowane wartości dla każdego z wyznaczonych czterech klastrów. Elementy skupienia pokazują natomiast które marki zostały zakwalifikowane do ktorego klastra. Możliwy jest też wgląd w rozkład wartości każdej z analizowanych cech w zależności od klastra.

Aglomeracja metodą Warda

W efekcie uruchomienia tej klasteryzacji zostanie nam przedstawione drzewo z rozgałęzieniami na kolejne klastry dla poszczególnych odległości pomiędzy analizowanymi rekordami.

Porównianie metod

Obydwie metody bazują na odległości w mierze euklidesowej i efektywnie różnica sprowadza się do formatu prezentacji danych oraz do faktu, iż w drugim przypadku sami możemy wybrać stosowny poziom drzewa dla którego chcemy przeprowadzić klasteryzację rekordów.

Porównajmy wyniki klasteryzacji zestawiając je w tabeli. Ponieważ dla k-średnich określiliśmy ilość skupień na cztery, natomiast diagram drzewa nie ma poziomu, po którego odcięciu otrzymalibyśmy cztery klastry, przetniemy drzewo na poziomie odległości równym 5, co da nam pięć klastrów

Nr klastra	k-średnie	drzewo m. Warda
1	Acura,	Acura,
	Buick,	Olds,
	Chrysler,	Chrysler,
	DOdge,	Dodge,
	Honda,	VW,
	Misub.,	Honda,
	Nissan,	Pontiac,
	Olds,	Nissan,
	Pontiac,	Mitsub.
	Saab,	
	Toyota,	
	VW,	
	Volvo	

Nr klastra	k-średnie	drzewo m. Warda
----------	----------	----------
2	Eagle	Audi,
		Mercedes,
		BMW,
		Saab,
		Volvo,
		Buick,
		Mazda,
		Toyota,
		Ford
----------	----------	----------
3	Audi,	Corvette,
	BMW,	Porsche
	Corvette	
	Ford,	
	Mazda,	
	Mercedes,	
	Porsche	
----------	----------	----------
4	Isuzu	Eagle
----------	----------	----------
5		Isuzu
----------	----------	----------

Chociaż klastry się od siebie trochę różnią, widać spore zbieżności i jeśliby odpowiednio rozdzielić rekordy z klastra nr 2 uzyskanego z drzewa pomiędzy klastry nr 1 i 3, to dla każdego klastra k-średnich moglibyśmy znaleźć identyczny klaster z drzewa.
Jednakże w obecnym kształcie klastry zostały utworzone odrobionę inaczej.

Analiza skupień dla All_cars.sta

Zbiór danych z Auta_all.sta jest tak duży, że ich ręczne porównywanie nie jest możliwe, natomiast przyglądając się drzewu utworzonemy metodą Warda można zauważyć, że odcinając dendrogram na

odległości ok 50.000 uzyskamy cztery skupienia, czyli dokładnie tyle samo, ile zadaliśmy metodzie k-średnich. Ponieważ i w tym przypadku metryką bazową dla obydwy metod klasteryzacji jest odległość euklidesowa, to skupienia te powinny sobie odpowiadać.