Rozpoznawanie: Klasteryzacja zbioru ofert sprzedaży mieszkania.

Paweł Szołtysek

Spis treści

Streszczenie

W ostatnim okresie ceny nieruchomości we Wrocławiu i okolicach znacznie się zwiększyły. W obliczu tego, zakup mieszkania coraz częściej traktowany jest jako inwestycja długoterminowa. Powoduje

to zwiększenie wykorzystywania matematycznych metod wyznaczania opłacalności takiego zakupu.
Niniejsza praca przedstawia wykorzystanie metod klasteryzacji do określania opłacalności zakupu nieruchomości. Na wstępie określono ogólny zbiór cech, które mają duży wpływ na ofertę sprzedaży, a następnie go zawężono. Wybrano losowo ponad sto ofert sprzedaży mieszkań na terenie aglomeracji wrocławskiej, po czym określono dla każdej z nich konkretne wartości cech. Do samej klasteryzacji wykorzystano autorską modyfikację metody DBSCAN.
Uzyskano zadowalające rezultaty - zostały określone takie nieruchomości, które na tle innych były faktycznie lepsze cenowo.

1 Wstęp

Po wstąpieniu Polski do Unii Europejskiej, sytuacja na lokalnych rynkach nieruchomości została bardzo zachwiana. Wraz z napływem zachodnioeuropejskiego kapitału, ceny mieszkań, domów i gruntów znacznie rosły, osiągając, a nawet przekraczając, poziom niemiecki czy francuski. Wobec nagłych wzrostów cen, wbrew prawu popyt-podaż, rósł też popyt na ten towar - wysoka obawa przed kolejną falą wzrostu nakręcała spiralę podwyżek.
Na pojedynczym, dużym serwisie ogłoszeniowym (których w Polsce jest kilkadziesiąt) przybywa setki, jeśli nie tysiące, ofert nieruchomości dziennie z terenu aglomeracji wrocławskiej. Wobec takiej ilości ofert, istotna jest wiedza o tym, które z nich są bardziej konkurencyjne względem innych - czyli charakteryzują się lepszym stosunkiem jakości nieruchomości do ceny.

2 Wygląd oferty sprzedaży nieruchomości

Standardową ofertę sprzedaży nieruchomości (mieszkania) można podzielić na kilka części:

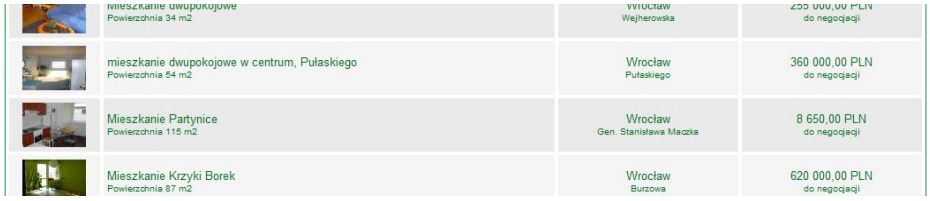
- dotycząca lokalizacji (ulica, okolica, komunikacja...),
- dotycząca budynku (ilość pięter, rok wybudowania...),
- dotycząca samego mieszkania (wielkość, ilość pokoi, wyposażenie...),
- dotycząca informacji niemierzalnych (miłe sąsiedztwo, ładny widok...).

Każda z nich ma wpływ na to, czy oferta jest atrakcyjna i powinna mieć swoje odzwierciedlenie w końcowej klasteryzacji ofert.

Rysunek 1: Lista ofert sprzedaży na jednym z serwisów ogłoszeniowych.

2.1 Ogłoszenia problematyczne

Podczas przeglądania ofert sprzedaży można zauważyć, że wiele z nich jest dalekich od doskonałości. Przykład takiego ogłoszenia widać na obrazku 2, gdzie sprzedawca źle wpisał cenę - zamiast łącznej, podał cenę za metr kwadratowy. W tym wypadku dokonanie naprawy jest proste (inteligencja pod-



Rysunek 2: Błędnie sformułowana oferta sprzedaży mieszkania.

powiada, że skoro w okolicy są w tej cenie jednostkowej mieszkania, to na pewno jest to błąd tego typu), ale taka sytuacja musi zostać znaleziona, zweryfikowana i zrobiona manualnie. Z tego też powodu przygotowanie zbioru ogłoszeń, które rozpatrujemy, jest trudniejsze niż w innych zadaniach klasteryzacji.

3 Wybór cech

Podczas doboru odpowiedniego zbioru, na którym będziemy pracowali, założymy, że zajmiemy się pewnym szczególnym, ale najbardziej rozpowszechnionym typem nieruchomości. Będą więc to lokale mieszkalne (tzn. trwale wydzielone fragmenty nieruchomości, które służą zaspokojeniu potrzeb bytowych osoby lub kilku osób), które pochodzą z rynku wtórnego (nie są sprze-

dawane przez deweloperów), oraz są mieszkaniami własnościowymi, bez zadłużenia. Ponadto ograniczymy się do terenu aglomeracji wrocławskiej, w celu lepszej pracy algorytmu.
Postawione ograniczenia pozwalają nam na znaczne uproszczenie całego procesu klasteryzacji.
Główne cechy, jakimi każda oferta się charakteryzuje, są przedstawione w poszczególnych podsekcjach.

3.1 Cechy lokalizacyjne

Do cech lokalizacyjnych zaliczamy:

miasto, którego oferta dotyczy,

dzielnica, w której się znajduje mieszkanie,

ulica, przy której znajduje się mieszkanie,

komunikacja miejska działająca w okolicy,

dojazd oraz odległość od centrum Wrocławia,

tereny zielone, które się znajdują w pobliżu mieszkania.

W procesie klasteryzacji nie będziemy brali pod uwagę pierwszych trzech atrybutów. Możemy założyć, że interesują nas wszystkie mieszkania w równym stopniu z badanego zbioru, a elementem, który decyduje o tym, czy jego lokalizacja jest dobra, pozostaje aspekt komunikacji miejskiej, dojazdu i obecności terenów zielonych.

3.2 Cechy budynku

Do głównych cech budynku zaliczymy:

rok budowy budynku,

materiał z którego budynek został wykonany,

ilość pięter budynku w pionie w którym mieszkanie się znajduje,

apartamentowiec - czy spełnia warunki do określania go tym mianem,

remonty - czy były wykonywane w ciągu ostatnich 5 lat generalne remonty.

Tutaj pod uwagę będziemy brali wszystkie cechy poza drugą. Powodem jest fakt, iż w małej ilości ofert znaleziono specyfikację jaką metodą budynek został wybudowany.

3.3 Cechy mieszkania

Do zasadniczych cech mieszkania zaliczymy:

wielkość mieszkania wyrażoną w metrach kwadratowych,

ilość pokoi,

ilość łazienek oraz toalet,

piętro na którym znajduje się mieszkanie,

wyposażenie które także podlega sprzedaży,

wysokość mieszkania w metrach,

wielkość czynszu wyrażona w złotych,

miejsce postojowe - czy jest, i czy spełnia normy garażu,

remonty - czy były wykonywane w ciągu ostatnich 5 lat generalne remonty.

cena mieszkania.

Tutaj wykorzystane zostały wszystkie cechy poza wysokością mieszkania oraz wielkości czynszu, z powodu analogicznego do tego w 3.2 - w zbyt małej ilości ofert występowały informacje na ten temat.

4 Przeprowadzenie klasteryzacji

Dane zostały zaczerpnięte z internetowych, ogólnodostępnych serwisów ogłoszeniowych, takich jak otodom, emieszkania czy domiporta. Każda oferta została zapisana w postaci wspomnianych wcześniej 15 cech, z których jedna odgrywa największą rolę - cena.
Po zebraniu ofert, wartości cech zostały znormalizowane, po czym przekazane do algorytmu klasteryzującego.
W celu przeprowadzenia klasteryzacji, zaimplementowany został w C++ algorytm DBSCAN. Jest to metoda prosta, bazująca na gęstości.
W celu znalezienia najlepszych parametrów dla niej, będziemy manewrować atrybutem ϵ.

4.1 Analiza danych wejściowych

Jak wspomniałem, wszystkie dane podległy normalizacji. Należy jednak powiedzieć o tym, jakie wartości obecnie prezentują poszczególne cechy i jak się to przekłada na liczby w świecie rzeczywistym.

Komunikacja miejska była wprowadzana jako wartość dyskretna z zakresu 0 - 5, w związku z czym różnica między stopniami wynosi 0,2.

Dojazd był określany jako wartość dyskretna z zakresu 0 - 6, różnica między stopniami wynosi 0,166.

Tereny zielone były określane jako wartość dyskretna z zakresu 0 - 4, różnica między stopniami wynosi 0,25.

Rok budowy określany był jako wartość dyskretna z zakresu 0-4, różnica między stopniami wynosi 0,25.

Ilość pięter maksymalnie wyniosła 11, przez co różnica między stopniami wynosi 0,0909.

Apartamentowiec jest wartością typu bool.

Remonty jest wartością typu bool.

Wielkość maksymalnie wyniosła 148m^2, w związku z czym 0,1 przekłada się na 1,48m^2.

Ilość pokoi w największym mieszkaniu wyniosła 6, różnica między stopniami to 0,166.

Ilość łazienek oraz toalet to 0 dla braku (nie wystąpiło), 0,5 dla jednej oraz 1 dla dwóch.

Piętro podobnie jak ilość pięter wyniosło maksymalnie 11, różnica między stopniami wynosi 0,0909.

Wyposażenie było określane jako wartość dyskretna z zakresu 0 - 4, różnica między stopniami wynosi 0,25.

Miejsce postojowe to 0 dla braku, 0,5 dla miejsca postojowego oraz 1 dla miejsca garażowego.

Remonty jest wartością typu bool.

Cena maksymalnie wyniosła 1080000 zł, w związku z czym 0,1 przekłada się na 108000 zł.

4.2 Tuning danych wejściowych

Jak widać w podsekcji 4.1, różne cechy różnie prezentowały swoje odniesienie w świecie rzeczywistym. Dla przykładu, mająca znacznie mniejszy wpływ informacja o ilości pięter w budynku miała większą wartość niż cena.
W celu uniknięcia takich ewenementów, każdej z cech przypiszemy wagę, którą podczas klasteryzacji będziemy próbowali poprawnie określić.
W trakcie badań, najlepsze i najbardziej zbliżone do realiów wagi okazały się:

Komunikacja miejska 0.5.

Dojazd 0.5.

Tereny zielone 0.25.

Rok budowy 1.

Ilość pięter 0.1.

Apartamentowiec 1.

Remonty 1.

Wielkość 1.

Ilość pokoi 0.5.

Ilość łazienek oraz toalet 0.1

Piętro 0.25

Wyposażenie 0.4.

Miejsce postojowe 0.1.

Remonty 1.

Cena 2.

4.3 Dwa podejścia do klasteryzacji metodą DBSCAN

W trakcie przygotowania do klasteryzacji okazało się, że można do niej podejść z dwóch różnych stron.

- **Dzielenie przestrzeni ofert mieszkań bez cen**
  Dokonując klasteryzacji samych mieszkań uzyskamy podział ofert na takie, które są względem siebie bardzo podobne. Następnie do takich podziałów dołożymy ceny, z jakimi oferty się pojawiły, które będziemy minimalizować, i dla każdej z klas wyznaczymy najlepszą ofertę.

- **Dzielenie przestrzeni ofert mieszkań z cenami**
  Biorąc pod uwagę także ceny podczas klasteryzacji, klasy będą zawierały oferty *standardowe*, to znaczy takie, które nie będą się różniły zdecydowanie między sobą wszystkimi cechami, z ceną włącznie. Wyszukiwanie atrakcyjnych ofert w takim wypadku sprowadzi się do oceniania tak zwanych szumów, czyli takich ofert, które nie zostały sklasyfikowane do żadnej klasy.

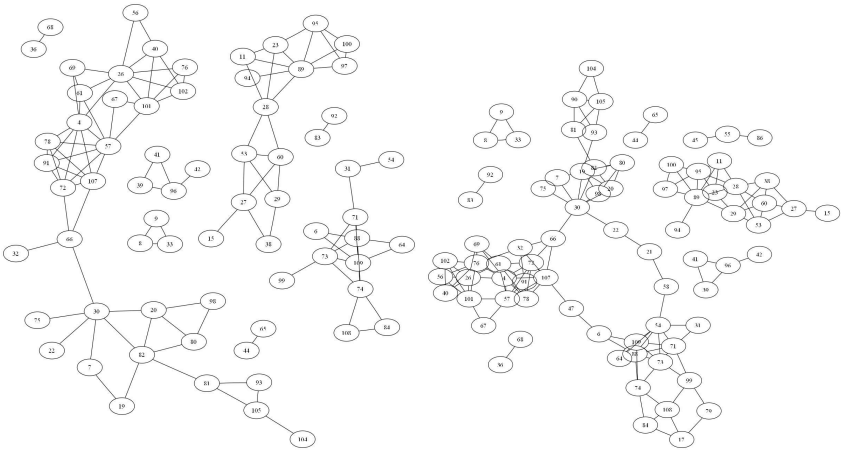
Ponadto, zostały wprowadzone pewne zmiany odnośnie określania bliskości dwóch ofert. Badania zostały przeprowadzone zarówno dla klasycznej formy metody DBSCAN, jak i dla takiej, w której ϵ określa maksymalną łączną różnicę wartości poszczególnych cech.
Jak więc widać, wprowadzono pewne autorskie urozmaicenia i rozszerzenia w stosunku do klasycznej metody DBSCAN.

5 Wyniki klasteryzacji

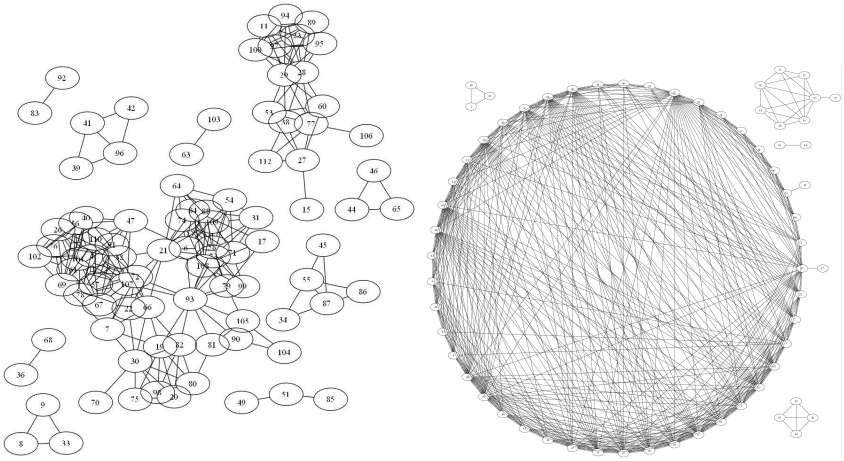
Przy obrazowaniu wyników klasteryzacji została wykorzystana aplikacja GVEdit 0.99 beta. Najczęściej używanymi silnikami były fdp oraz circo.

5.1 Przypadek *z cenami*

Dla $\epsilon = 0.1$ znaleziono tylko dwa dopasowania (tj. wszystkie klasy poza dwoma składały się z jednego elementu).
Dla $\epsilon = 0.2$ znaleziono tylko 20 dopasowań.
Dla $\epsilon = 0.25$ wystąpiło 65 niesklasowanych ofert.
Dla $\epsilon = 0.3$ - 41 niesklasowanych ofert.
Dla $\epsilon = 0.4$ - 39 niesklasowanych ofert.
Dla $\epsilon = 0.49999$ - 32 niesklasowanych ofert.
Dla $\epsilon = 0.5$ - 11 niesklasowanych ofert.

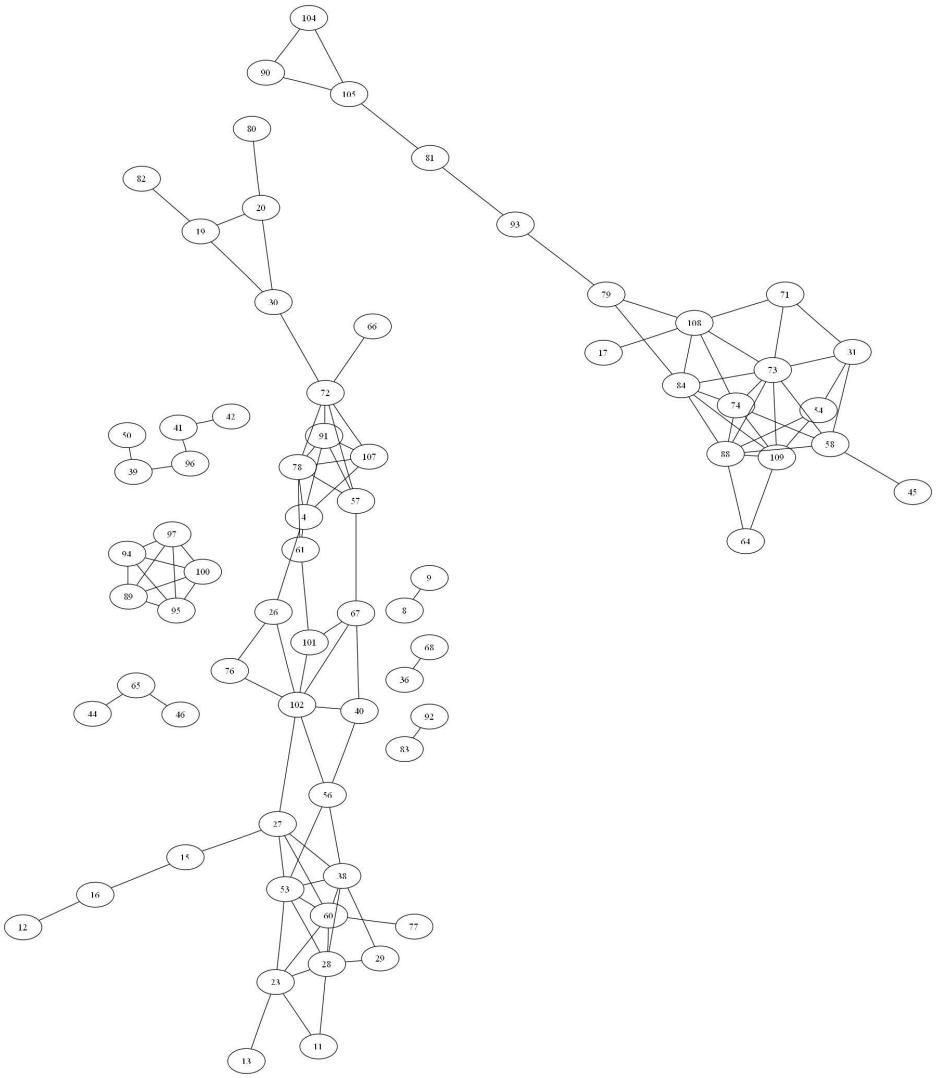


Rysunek 3: $\epsilon = 0.25$, $\epsilon = 0.3$



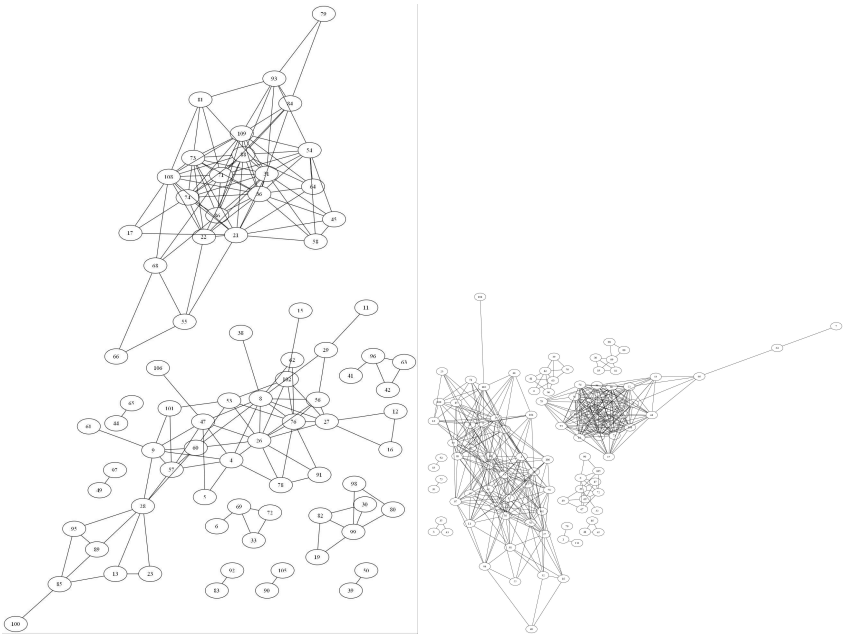
Rysunek 4: $\epsilon = 0.4$, $\epsilon = 0.5$

Uzyskane wyniki nie zaskakują. Ilości podobnych ofert zgodnie z oczekiwaniami są skokowe względem wartości ϵ.
Jest jasne, że nie można wykorzystać wyników dokonanej klasteryzacji w sposób poprawny. Dla $\epsilon = 0.4$ niesklasyfikowanych ofert pozostaje zbyt dużo, natomiast dla $\epsilon = 0.5$ zdecydowana ich większość tworzy jedną, bardzo dużą klasę (widać to dobrze na rysunku 4). Należy jeszcze dodać, że różnica ceny w tej klasie pomiędzy ofertami wynieść $2 * \epsilon = 1080000$ zł.
Dla drugiej metody obliczania podobieństwa dwóch ofert, najlepsze wyniki otrzymano dla $\epsilon = 1$, co przedstawia rysunek 5. Otrzymany wynik jednak nadal pozostawiał wiele do życzenia, z powodu ilości obiektów które tworzyły samodzielnie klasy (60).



Rysunek 5: $\epsilon = 1$

Wobec tego należy skorzystać z wag dla cech (przede wszystkim ceny). Do tego celu użyjemy wag określonych w 4.2.
Dla $\epsilon = 0.1$ 80 elementów tworzyło szum.
Dla $\epsilon = 0.15$ wartość ta wyniosła 53.
Dla $\epsilon = 0.2$, 28 elementów było niesklasyfikowanych.
Dla $\epsilon = 0.25$, 18 elementów było niesklasyfikowanych.
Należy zauważyć przy tym, ze dla $\epsilon = 0.25$ ilość krawędzi pomiędzy elementami była prawie dziesięciokrotnie większa niż dla $\epsilon = 0.2$. Dla tych danych wydaje się, że $\epsilon = 0.2$ jest najlepszą wartością.



Rysunek 6: $\epsilon = 0.15$, $\epsilon = 0.2$

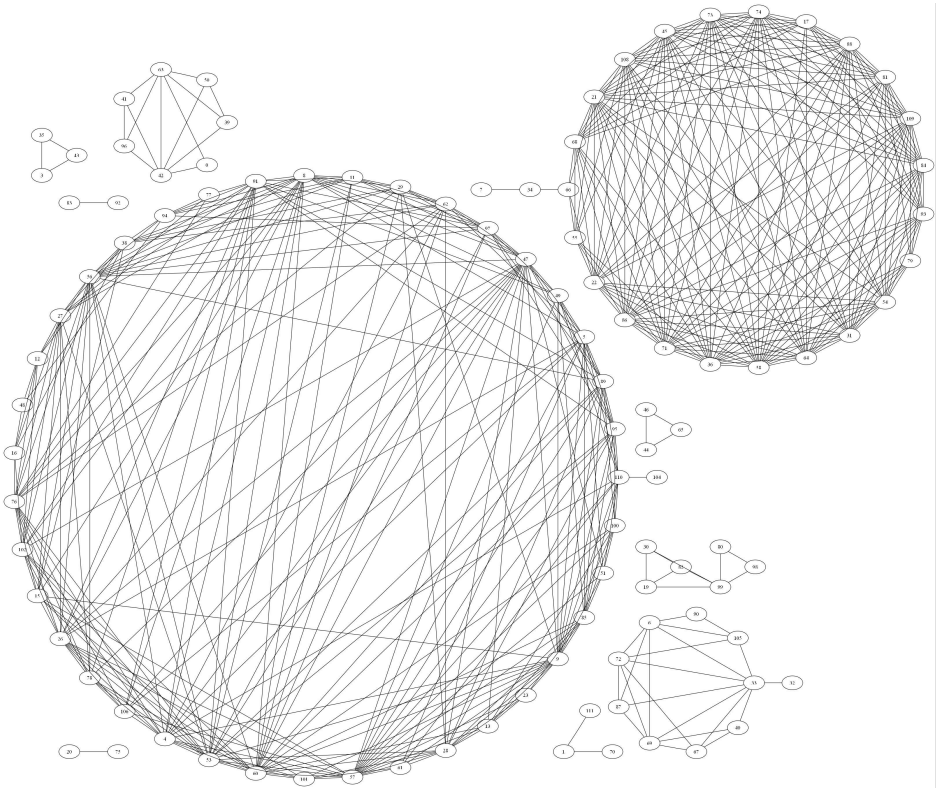
Sprawdźmy też użycie wag przy metryce całościowej.
Dla $\epsilon = 0.5$ - 52 elementy tworzą szum.
Dla $\epsilon = 0.625$ - 38 elementów tworzy szum.
Dla większych wartości ilość szumu nie spada korzystnie względem tworzenia się nowych połączeń.

Wydaje się więc, że dla tego przypadku optymalnym rozwiązaniem jest wykorzystanie wag, oraz określenie $\epsilon = 0.2$ dla miary osobnej dla każdej cechy. Przeanalizujmy dokładnie ten przypadek.

5.1.1 Analiza najlepszego przypadku

Dla wag znajdujących się w 4.2, $\epsilon = 0.2$ oraz miary osobnej dla każdej cechy, graf bliskości ofert został przedstawiony na rysunku 8. Interesujące nas rekordy stanowią szum który powstał po takiej klasteryzacji. Są to odpowiednio:

Rysunek 7: $\epsilon = 0.5$, $\epsilon = 0.625$



Rysunek 8: $\epsilon = 0.2$

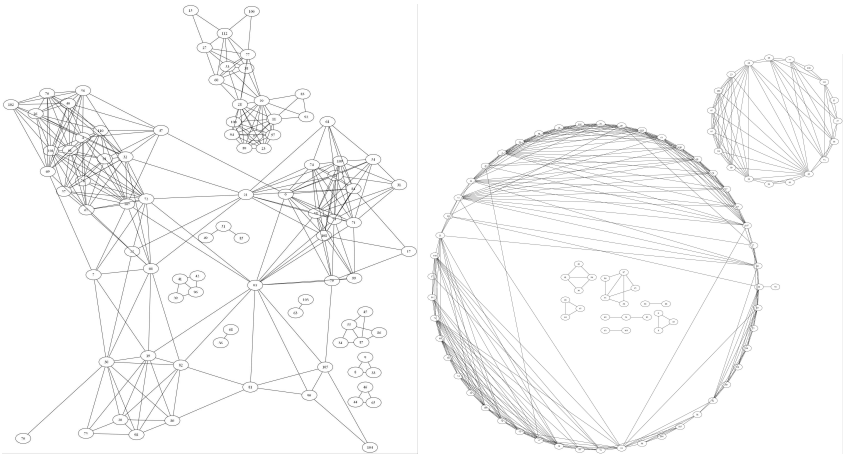
2	Kiełczowska	5	3	4	0	1	0	0	110	4	1	0	2	2	0	780000
10	Obr. Poczty Gd.	3	3	4	4	4	0	0	50	2	1	1	2	2	0	340000
14	Lipowa	2	2	4	4	2	0	0	39	2	1	2	0	2	0	285000
18	Gajowicka	4	5	2	2	11	0	0	44	2	1	11	2	0	0	290000
24	Szkolna	2	2	4	4	2	0	0	55	2	1	1	0	1	0	265000
25	Miernicza	5	6	1	0	4	0	0	68	2	1	3	2	0	1	469000
37	Popowicka	4	4	4	2	4	0	0	41	2	2	2	2	0	0	265000
43	Zemska	4	3	3	2	4	0	0	50	2	1	2	4	0	1	348000
52	Główna	5	3	3	4	5	0	0	53	2	1	1	0	0	0	280000
59	Dokerska	5	3	2	2	9	0	0	36	2	1	4	2	0	0	265000
65	Wejherowska	3	4	4	2	10	0	0	34	2	1	4	2	0	0	245000
70	Poranna	3	4	4	3	4	0	0	57	2	1	0	2	0	0	420000
75	Orzechowa	5	5	3	2	10	0	0	47	2	1	6	2	0	0	320000
87	Jabłeczna	5	5	3	2	10	0	1	54	2	2	0	3	0	0	320000
92	Gradowa	4	3	2	4	3	0	0	66	3	1	3	2	0	0	495000
96	Ułańska	3	3	3	4	2	0	0	57	2	1	0	0	1	0	398000
97	Saperów	4	4	2	2	4	0	0	57	3	1	4	3	0	1	429000
99	Komandorska	5	5	2	1	4	0	0	27	1	1	0	3	0	0	267000
101	Motylkowa	3	4	3	4	2	0	0	42	2	1	1	0	1	0	304590
102	Vivaldiego	2	2	3	4	3	0	0	47	2	1	1	3	0	0	370000
103	Chabrowa	2	2	4	4	3	0	0	64	2	1	1	3	0	0	490000
105	Komuny Paryskiej	4	6	0	4	6	0	0	47	2	1	5	3	0	0	315000
106	Swobodna	4	5	0	3	6	0	1	51	2	1	5	3	0	0	350000
107	Gaj	5	5	3	4	4	0	0	62	3	1	4	3	1	0	449000
109	Buska	5	5	2	2	10	0	0	58	3	1	8	2	0	0	340000
110	Drukarska	4	5	3	2	10	0	0	38	2	1	4	2	0	0	295000
111	Armii Krajowej	5	5	3	4	3	0	0	46	3	1	3	2	0	0	320000
112	Rodzinna	4	3	2	0	3	0	0	122	3	1	2	0	1	0	850000

gdzie odpowiednie kolumny oznaczają wcześniej opisane cechy. Wśród tych wyników powinniśmy poszukiwać ofert, które uważamy za ciekawe.

5.2 Przypadek *bez cen*

Przeanalizujemy teraz wspomniany wcześniej przypadek, gdy przy klasteryzacji nie bierzemy pod uwagę cen. Będziemy tutaj dążyć do zbalansowanego podziału zbioru mieszkań na klasy, w których będziemy mieli oferty w pewnym stopniu podobne do siebie. W tych klastrach będziemy z kolei dążyli do minimalizacji cen. W ten sposób otrzymamy zbiór ofert w których możemy poszukiwać promocji cenowych.
Zaczniemy podobnie jak wyżej, czyli bez wag oraz z metrykami jak w stan-

dardowej metodzie DBSCAN.
Przy $\epsilon = 0.3$, 51 elementów tworzy szum.
Przy $\epsilon = 0.4$, 38 elementów tworzy szum.
Przy $\epsilon = 0.5$ 11 elementów tworzy szum, ale z uwagą jak wyżej - w pojedynczym klastrze mogą znajdować się oferty o diametralnie różnych właściwościach.
W tym wypadku jest 322 połączeń między ofertami, zostały utworzone dwa



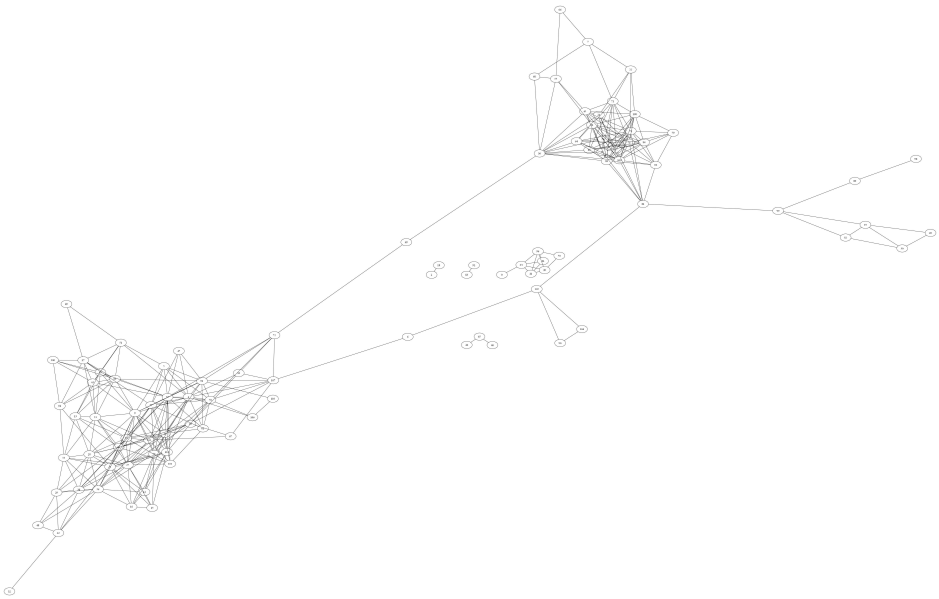
Rysunek 9: $\epsilon = 0.4$

główne klastry, które można podzielić (usuwając niewielką liczbę połączeń) wyraźnie na pięć dużych podklastrów.
Po zmianie metryki, tzn. oznaczeniu ϵ jako całościowej granicy, otrzymujemy następujące wyniki.
Dla $\epsilon = 1.1$, 46 elementów tworzy szum.
Już na tym etapie jednak uzyskane wyniki tworzą jeden, wysoce spójny klaster. Nie ma więc sensu kontynuować badań.
Wprowadźmy więc wagi, i użyjmy standardowej metryki.
Dla $\epsilon = 0.1$, 61 elementów tworzy szum. Warty zauważenia jest fakt, że zostało utworzonych wiele klas które mają w sobie kilka obiektów.
Dla $\epsilon = 0.125$, 44 elementów tworzy szum. Powstał wyraźny podział na dwa klastry, i trzeci mniejszy.
Dla $\epsilon = 0.15$, 38 elementów tworzy szum. Klastry jednak rozpoczęły się już jednak zespalać.

Pozostając przy wagach i ustalając drugą metrykę, uzyskujemy wyniki:
Przy $\epsilon = 0.4$, 44 oferty tworzą szum przy 197 połączeniach.
Przy $\epsilon = 0.45$, 36 oferty tworzą szum przy 275 połączeniach.

Rysunek 10: $\epsilon = 0.125$

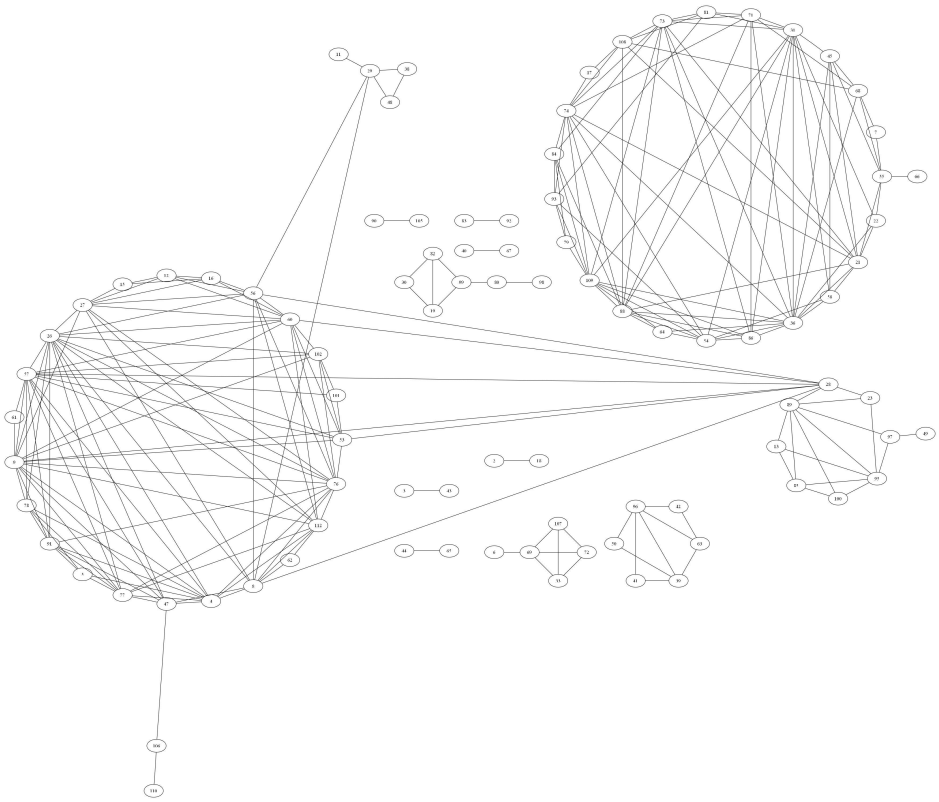
Przy $\epsilon = 0.5$, 33 oferty tworzą szum. Mamy 371 połączeń i dwa klastry.



Rysunek 11: $\epsilon = 0.5$

5.2.1 Analiza najlepszego przypadku

Najlepsze efekty, czyli największą ilość klastrów możemy zaobserwować w przypadku pierwszym oraz trzecim, tj. tam, gdzie była brana pod uwagę standardowa metryka.
Będziemy rozpatrywali przypadek z wagami, i $\epsilon = 0.125$.



Rysunek 12: $\epsilon = 0.125$

Uzyskane wyniki muszą zostać teraz poddane obróbce w taki sposób, aby z każdego klastra uzyskać te oferty, których cena była najmniejsza. Będą to m.in.

24	Szkolna	2	2	4	4	2	0	0	55	2	1	1	0	1	0	265000
40	Fiołkowa	5	4	2	2	3	0	0	45	2	1	3	2	0	1	255000
70	Poranna	3	4	4	3	4	0	0	57	2	1	0	2	0	0	420000

gdzie odpowiednie kolumny oznaczają wcześniej opisane cechy.
Należy zauważyć, że dwa z nich znajdują się także w tabeli w 5.1.1.

6 Podsumowanie

Praca miała na celu przeprowadzenie eksperymentu, który polegał na zastosowaniu czystomatematycznych metod klasteryzacji zbioru do realnie istniejących ofert sprzedaży lokali mieszkalnych. Z powodów oprogramowania generującego graficzne postacie grafów przedstawiających klastry, został ograniczony zbiór który podlegał klasteryzacji do 113 rekordów.

Po przeprowadzeniu badań przeanalizowano też zachowanie algorytmów dla zmniejszonej ilości ofert. Dla 69 rekordów, żadna z propozycji zastosowania metody DBSCAN nie była w stanie wygenerować więcej niż dwóch dużych klastrów przy akceptowalnej wielkości szumów. Można z tego wywnioskować, że przy dużej próbie, osiągniętoby lepsze wyniki w sensie ilości i różnorodności klas wynikowych (co z kolei ma duże znaczenie przy podejściu przedstawionym w 5.2).
Ogólnie jednak można powiedzieć, że uzyskane wyniki są akceptowalne, a różne strategie podejścia do problemu potrafią dawać podobne rezultaty.

Literatura

[1] Martin Ester, Hans-Peter Kriegel, Joerg Sander, Xiaowei Xu: *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, http://omen.cs.uni-magdeburg.de/itikmd/fileadmin/downloads/SS07/DMCE/DBScan.pdf

[2] Adam Lessnau: *Klasteryzacja*, http://www.fizyka.umk.pl/∼duch/zajecia/05SemMagInf/02Klasteryzacja.pdf

[3] Nguyen Hung Son: *Clustering. Efektywne metody grupowania danych*, http://www.mimuw.edu.pl/∼son/datamining/materials/w9_cluster.pdf