Wykorzystanie algorytmów rojowych do klasteryzacji zbioru ofert sprzedaży mieszkań Algorytmy genetyczne - Projekt

Maciej Kupczak, Paweł Szołtysek

Spis treści

1 Wstęp

Na każdym z dużych serwisów ogłoszeniowych dotyczących nieruchomości (których w Polsce jest kilkadziesiąt), dziennie przybywa setki, jeśli nie tysiące ofert nieruchomości z terenu aglomeracji wrocławskiej. Wobec takiej ilości,

istotną wydaje się wiedza o tym, które z ofert są *lepsze* względem innych - czyli charakteryzują się lepszym stosunkiem jakości nieruchomości do ceny. Problem jest ważny przede wszystkim dla osób, które poszukują mieszkań na własną rękę - dość często takie osoby nie znają rynku, nie wiedzą jak powinny ocenić daną ofertę.

Oczywiście, zwykle nie można wprost określić, że dana oferta jest dobra. Już we wcześniejszych pracach [8], [9] autor zauważył, że poruszony problem można w odpowiedni sposób rozwiązać korzystając z techniki klasteryzacji ofert. Pomysł wykorzystaniu klasteryzacji opiera się na grupowaniu ofert które uznajemy za podobne pod względem ogólnym lub ofertowym (zależnie od podejścia) i odpowiednio szukania ofert w danych klastrach lub szumie. Ponadto, w [9] zostało przedstawione wykorzystanie algorytmu DBSCAN do realizacji tego zadania. Ponieważ praca zakończyła się sukcesem, mając na względzie rosnącą ilość ofert które pojawiają się w serwisach ogłoszeniowych autorzy wychodzą z założenia, że należy zastosować bardziej efektywny, a przede wszystkim szybszy sposób klasteryzacji założonego zbioru.

Celem projektu jest zaproponowanie i implementacja algorytmu klasteryzującego zbiór ofert sprzedaży mieszkań na podstawie algorytmu rojowego. Prace będą prowadzone w oparciu o wcześniejsze prace dotyczące tematu klasteryzacji danych dotyczących sprzedaży mieszkań. Finałowym efektem prac będzie porównanie opracowanego algorytmu z wcześniej wykorzystanym prostym algorytmem DBSCAN.

Praca jest podzielona na dwie części: pierwsza, teoretyczna, (sekcje 2-5) opisuje krótko istotę algorytmów PSO, ich zastosowanie do klasteryzacji oraz wymienia cechy ofert, pod względem których będzie dokonywana klasteryzacja. Druga, praktyczna (sekcje 6-8) skupia się na implementacji oraz testowaniu zaproponowanego algorytmu, a także porównaniu efektów jego działania do zaimplementowanego wcześniej algorytmu DBSCAN.

2 Oferty sprzedaży mieszkań

Na opłacalność oferty sprzedaży mieszkania ma wpływ wiele różnych czynników, od lokalizacyjnych (ulica, komunikacja publiczna, zieleń, itp), poprzez właściwości budynku (rok budowy, standard, ilość pięter, itp) i lokalu (wielkość, ilość pokoi, stan prawny, wyposażenie, itp) aż po te dotyczące informacji niemierzalnych (miłe sąsiedztwo, jasna kuchnia, itp), a o niektórych z nich nie można wprost powiedzieć, że wprost wpływają pozytywnie lub negatywnie na ofertę. Różni kupujący będą bowiem różnie interpretowali propozycje - samotna osoba nie będzie chciała prawdopodobnie kupić dużego, wielopokojowego mieszkania przez fakt wysokich kosztów miesięcznego utrzymania takiej nieruchomości.

Na potrzeby implementacji niniejszego projektu wykorzystany zostanie taki sam zbiór danych jak we wcześniejszej pracy [9]. Pozwoli to na bezpo-

średnie porównanie omawianych algorytmów w kontekście ich skuteczności oraz efektywności.

Należy jednak podkreślić fakt zauważony już w powyższej pracy: nie jest możliwe całkowicie automatyczne zebranie informacji, które są wymagane do pracy algorytmu, a które są publicznie dostępne w serwisach ogłoszeniowych. Niektóre z nich są przedstawiane w odpowiedni sposób zgodnie z nurtem semantyki w sieci (np. powierzchnie czy lokalizacje w serwisie otodom), jednak nadal wiele informacji które z punktu widzenia kupującego są bardzo ważne są przedstawiane w formie tekstu, co znacznie utrudnia automatyczne *rozumienie* informacji w ofercie. Ponadto niezwykle często zdarza się, że różne informacje są błędnie wpisywane przez ogłoszeniodawców (Przykład: parametr cena jest podawany przez niektórych jako cena łączna, a przez niektórych jako cena za metr kwadratowy) - jest to proste do zweryfikowania przez człowieka, ale dla maszyny jest to prawie niemożliwe.

Problem tworzenia zbioru ofert nie będzie w ninejszej pracy omawiany - zostanie założone, iż użytkownik systemu jest już w posiadaniu odpowiedniego zbioru przystosowanego do pracy algorytmu. Trzeba jednak powiedzieć, że taki zbiór można uzyskać mając dostęp do bazy danych serwisu ogłoszeniowego, lub korzystając w *WebCrawlera* taką bazę na własną rękę stworzyć. W przypadku danych opisowych, lub podawanych niewprost należy wykorzystać różne dostepne na rynku rozwiązania typu data mining, które pozwolą na przekształcenie do użytecznej formy niezrozumiałych zapisów stworzonych przez użytkownika.

Ponadto można zauważyć, że w podobnych zastosowaniach ale do innych (prostszych) zbiorów etap ten będzie nieco uproszczony poprzez lepszą organizację ogłoszeń. Przykładem jest tu witryna autogielda, gdzie można znaleźć w sposób stricte semantyczny wpisane wszystkie podstawowe dane o samochodzie.

3 Algorytmy Particle Swarm Optimization

Idea algorytmu rojowego pochodzi z pracy [3] z 1995 roku - jest to więc dość młode zagadnienie. Od tej pory powstało wiele wariacji sposobów rozwiązań problemów związanych z tą ideą.

3.1 Wstęp do algorytmu PSO

Zachowanie pojedynczej mrówki, pszczoły czy termitu jest zwykle bardzo proste. Jednak o sile tego gatunku decyduje nie jednostka, jednak cała społeczność.

Powyższa obserwacja stanowi podstawę algorytmów w dziedzinie *Swarm intelligence*. Po pojawieniu się pierwszych prac wykorzystujących zachowanie owadów, wielu naukowców rozpoczęło obserwację naturalnego zachowa-

nia zwierząt i przekładanie sposobów radzenia sobie z problemami w realnym świecie na problemy dotyczące np. optymalizacji.

Ogólnie można więc powiedzieć, że takie algorytmy będą bazowały na zbiorze prostych osobników, wykonujących określone czynności, które będą ze sobą się w odpowiedni sposób komunikowały. Tego typu algorytmy są zwykle zaprojektowane w taki sposób, że nie występuje narzucone z góry sterowanie pracą wszystkich osobników, jednak dzięki założonym regułom całość osiąga założony cel.

W pracy [1] wyróżniono kilka zasadniczych cech zachowania, które są zakorzenione w naturze, a z którego korzystają algorytmy oparte na *Swarm Intelligence*:

Jednorodność każdy element ma z góry określony model zachowań, nie występuje stały lider

Lokalność tylko najbliższe elementy mają wpływ na zachowanie każdego elementu

Unikanie kolizji z elementami w okolicy

Dostosowanie prędkości z elementami które są w okolicy

Centrowanie stada próba trzymania się w odpowiedniej bliskości do innych elementów

3.2 Algorytm PSO

W algorytmach roju cząstek, populacja elementów jest inicjalizowana z losowymi pozycjami X_ai i prędkościami V_ai każdej z cząstek $a \in 1, 2, ..., A$ każdego z wymiarów $i \in 1, 2, ..., n$, oraz funkcja f.

W każdej z kolejnych iteracji algorytmu wylicza się nowe prędkości i pozycje każego z elementów. Równanie dla wymiaru d można przedstawić jako

$$V_{id}(t+1) = \omega V_{id}(t) + C_1\phi_1(P_{lid} - X_{id}(t)) + C_2\phi_2(P_{gd} - X_{id}(t))$$
$$X_{id}(t+1) = X_{id}(t) + V_{id}(t+1)$$

gdzie:

- $0 < \phi_1, \phi_2 < \phi_{max}$ - losowe wartości
- C_1, C_2 - stałe przyśpieszenia
- ω - współczynnik bezwładności
- P_{li} - najlepsze lokalne rozwiązanie (znalezione przez element i)
- P_g - najlepsze globalne rozwiązanie (znalezione przez dowolny element)

Oczekujemy, że po zakończeniu pracy algorytmu (warunek stopu - ilość wykonanych iteracji) większość elementów będzie w bliskiej okolicy globalnego optimum przestrzeni poszukiwań.

Listing 1: Algorytm PSO

```
while t<T do
  for a=1 to A do
    Wyznacz f(Xi(t))
    Uaktualnij Pli(t) oraz Pg(t)
    Wyznacz ponownie prędkość oraz położenie
```

4 Klasteryzacja oparta o algorytm PSO

Ponieważ klasteryzacja danych to szczególny przypadek problemu optymalizacji, tylko kwestią czasu było utworzenie algorytmu opartego o pomysł PSO który służyłby do klasteryzacji. Różne samodzielne algorytmy zostały przedstawione m.in w [7], [10] czy [1]. Jednocześnie jakość klasteryzacji takiego podejścia do klasteryzacji została zbadana już np. w zadaniu klasteryzacji dokumentów [2] czy obrazów [6]. Ponadto zostało pokazane, że istnieją przypadki, w których PSO pracuje znacznie lepiej niż popularne metody takie jak K-Means czy FCM.

Wobec powyższego, niniejsza praca prezentuję zastosowanie algorytmu PSO w kolejnym zastosowaniu - do klasteryzacji ofert, w szczególności ofert mieszkaniowych. Już w sekcji 1 napisano o sensie klasteryzacji takiego zbioru. W tym rozdziale zostanie przedstawiony zastosowany przez nas algorytm.

Poniższy algorytm stanowi zaproponowaną w [1] modyfikację klasycznego algorytmu klasteryzującego opartego na PSO (zaproponowanego z kolei w [7]) zwaną MEPSO. Wprowadza on z punktu widzenia prezentowanego tutaj zastosowania znaczne usprawnienie - ilość klas nie jest parametrem wejściowym, a jest określona dynamicznie przez algorytm. Jednocześnie jest to podstawowe założenie przedstawionego zastosowania (w przeciwnym wypadku należałoby korzystać ze specjalnych algorytmów szacujących ilość klas).

Listing 2: Algorytm MEPSO

```
for t=1 to T do
  if t<T then
        for a=1 to A do
              if f(Xi(t))>f(Xi(t-1)) then
                    bi=bi+1;
              Uaktualnij Pli
              if f(Pli)>f(Pg) then
                    Dodaj do kandydatów Pli
        Oblicz b każdego kandydata
```

```
10          Uaktualnij Pg kandydatem z bmax
11      else
12          Uaktualnij Pg elementem z najwyższą wartością f
```

5 Wybór cech

W pracy wykorzystany sprawdzony już w [8] zbiór cech określających ofertę mieszkaniową.

Dla przypomnienia, skupimy się na najbardziej popularnych nieruchomościach - lokalach mieszkalnych, które pochodzą z rynku wtórnego, oraz są mieszkaniami własnościowymi. Ograniczymy się do lokali nie posiadających zadłużenia, które są zlokalizowane na terenie aglomeracji wrocławskiej.

Takie ograniczenia pozwalają na znaczne uproszczenie całego procesu klasteryzacji, jednak nie wpływają na sposób jego działania i wnioski z pracy algorytmu.

Główne cechy, jakimi każda oferta się charakteryzuje, są przedstawione w poszczególnych podsekcjach.

5.1 Cechy lokalizacyjne

Do cech lokalizacyjnych zaliczamy:

komunikacja miejska działająca w okolicy,

dojazd oraz odległość od centrum Wrocławia,

tereny zielone, które się znajdują w pobliżu mieszkania.

Nie będziemy brali pod uwagę takich oczywistych cech jak ulica, przy której znajduje się nieruchomość - 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.

5.2 Cechy budynku

Do głównych cech budynku zaliczymy:

rok budowy budynku,

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.

5.3 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,

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

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

cena mieszkania.

6 Przeprowadzenie klasteryzacji

Klasteryzacja zostanie przeprowadzona dla takich samych danych jak w pracy [9]. Podobnie też zadanie klasteryzacji będzie wykonane korzystając z dwóch strategii wyszukiwania ofert.

- **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 najlepsze oferty.

- **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.

6.1 Implementacja algorytmu

7 Testowanie algorytmu

8 Porównanie algorytmów DBSCAN i PSO

9 Podsumowanie

Literatura

[1] Abraham A., Das S., Roy S. *Swarm Intelligence Algorithms for Data Clustering*, Oded Maimon and Lior Rokach (Eds.), Springer Verlag, Germany, pp. 279-313, 2007.

[2] Cui X., Potok T., Palathingal P. *Document Clustering using Particle Swarm Optimization*, IEEE Swarm Intellience Symposium, The Westin Pasadena, Pasadena, California, 2005.

[3] Eberhart, R. C., Kennedy, J. A. *New optimizer using particle swarm theory* Proceedings of the Sixth International Symposium on Micromachine and Human Science, Nagoya, Japan, pp. 39-43, 1995.

[4] Kennedy, J. *Stereotyping: improving particle swarm performance with cluster analysis* Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2000), San Diego, CA, pp. 1507-1512, 2000.

[5] Omran M. *Particle Swarm Optimization Method for Patter Recognition and Image Processing* rozprawa doktorska, University of Pretoria, 2005.

[6] Omran M., Engelbrecht A.P., Salman A. *Particle Swarm Optimization Method for Image Clustering* International Journal of Pattern Recognition and Artificial Intelligence, 19(3), pp. 297-322, 2005.

[7] Omran M., Salman A. and Engelbrecht A.P. *Image Classification using Particle Swarm Optimization* Conference on Simulated Evolution and Learning, volume 1, pp. 370 - 374, , 2002.

[8] Szołtysek P. *Rozpoznawanie: Klasteryzacja zbioru ofert sprzedaży mieszkania* Project Report, Wrocław University of Technology, 2007.

[9] Szołtysek P. *Usage of 'dbscan' clusterization method to the set of housing sale offers* 7th. Students' Scientific Conference: Man - Civilization - Future. Papers of KNS 2009.

[10] Van der Merwe D. W., Engelbrecht A. P. *Data clustering using particle swarm optimization* Proceedings of IEEE Congress on Evolutionary Computation 2003 (CEC 2003), Canbella, Australia. pp. 215-220, 2003