clusterization, real estate market, housing sale offers, data mining

Paweł SZOŁTYSEK*

USAGE OF 'DBSCAN' CLUSTERIZATION METHOD TO THE SET OF HOUSING SALE OFFERS

In the last years on the Polish real estate market prices of the housing were increasing pretty much. Because of that, buying a house is more often considered as an investment – and mathematic methods of evaluation of profitability might be useful for the investors.

This work presents usage of DBSCAN clustering method to the set of housing sale offers, which will result in a list of best offers from the initial set.

Whole process of preparing the data, determining the attributes and implementation of the algorithm has been presented. To the standard DBSCAN procedure, different method of calculating the measure has been proposed. Finally, basing on randomly chosen 113 offers from Wrocław agglomeration area, clusterization has been performed. Results gained from presented method were (compared to general situation at the market) offers worth consideration.

1. MOTIVATION

In the year 2004, when Poland joined the European Union, real estates were pretty cheap if compared to West European prices. Therefore, investors started to buy houses and flats, destabilizing Polish market of real estates, and making prices go higher and higher, reaching, or even surpassing German or French level. This process was even strengthen by Polish inhabitants, who were afraid of high prices, were buying them faster, making sellers demand more cash than before.

There are dozens of quite big Polish websites with housing offers, and each one is registering hundreds of them just from Wrocław agglomeration. With that number of them, the knowledge of current real estate market is crucial (because prices are changing quickly too), so if we want to buy a house, we would like to have a tool which would tell us which offers from specified set of them are favorable.

* Wrocław University of Technology, Wybrzeże Wyspiańskiego St. 27, 50-327 Wrocław, „ESTYMATOR", tutor: Krzysztof Brzostowski

2. HOUSING SALE OFFERS

Generally, offer can be divided into several parts:

- Localization (street, access to parks, public transport, …)
- Building (year of build, number of floors, …)
- House (number of rooms, size, equipment, …)
- Unmeasurable (nice neighbourhood, pretty view, …)

Each of them has influence whether an offer is attractive and should be taken into consideration while constructing final set of offers.

3. PREPARING CLUSTERIZATION

3.1. GATHERING DATA

At the beginning, it's needed to prepare the data which will be used to perform clusterization. As stated above, there are many websites which are providing offers. However, automatic processing cannot be easily used in this case. The data are usually not sorted semantically, or even if, they happen to be wrongly filled. Image 1 shows an example – someone has put price for one sq.m. instead of total price for the house. The solution is easy for human, but automatic processing usually fails in this case.

Image 1. Wrongly created offer of housing

Of course, it's necessary to focus on one type of real estate – mixing them might influence wrong clusterization (150 sq.m. house is different than 150 sq.m. flat). This work focuses on the second-hand flats, which rightfully belong to seller and are not on debt. The set of offers will be bounded to Wrocław agglomeration. As a result, clusterization process is easier, but still complies to any other situation.

In this work set of 113 offers was taken. By a coincidence, they are from June 2008, when the prices reached maximum [6].

3.2. SET OF ATTRIBUTES

In the next step, semantic data from the offers are extracted. Final set of attributes for each offer is as following.

- Localization - access to public means of transport; distance to the city centre; green areas around the house.
  City, street etc. are not considered here – this work is prepared only for the Wrocław agglomeration, so distance attribute provides all needed data.
- Building - year of build; number of floors; Is the building an apartment?; Were there any general renovation performed within last 5 years?.
  The material which was used to build the house is important too, but most offers doesn't carry such information (though it's easy to predict that – for example, most buildings taller than four floors and builded between 1960 an 1990 in Wroclaw were made from precast concrete slabs).
- House - size of house; number of rooms; number of bathrooms and toilets; floor on which flat is located; equipment which is added to the flat; Is there a parking place? Does it handle the garage requirements?; Were there any general renovation performed within last 5 years?.
  There's one important factor which is not included here – monthly fee. It's the minority of reports, which possess such information.
- Price.

All in all, each offer has been described using 15 attributes, where price is the most important. All values of attributes were normalized, analyzed, tuned (to make the process more valuable) and finally directed to the clusterization algorithm.

4. CLUSTERIZATION PROCESS

In clusterization process, well-known DBSCAN algorithm has been used. It’s easy and intuitive method, based on the density of given set. To find the best ε parameter, size of noise will be checked. The algorithm has been implemented using C++.

4.1. DIFFERENT STRATEGIES OF CLUSTERIZATION

While preparing the algorithm, it appeared that the clusterization can be performed using two different approaches.

- Clustering the offers without prices – in this case, houses will be clustered (so one class will respond to one type of them). In the next step prices will be added, and for each class, offer with the lowest price will be chosen.
- Clustering the offers with prices – in this strategy, classes will contain usual offers (which will not differ much in any of attributes, including price). The best offers will be found in the noise.

Some improvements to the DBSCAN method were implemented, mostly related to used measure. For each case, algorithm was running twice – once for standard method of measuring ε, and once where ε means total difference between offer attributes.

5. CLUSTERIZATION RESULTS

To display the results the GVEdit 0.99 beta was used, with fdp and circo engines.

5.1. CASE WITH PRICES

Table 1 presents the results for the clusterization in the case with prices.

Table 1. Results of clusterization with prices

ε	Number of unclassified offers standard DBSCAN procedure	Number of unclassified offers modified DBSCAN procedure
0.1	80	-
0.15	53	-
0.2	**28**	-
0.25	18	-
0.5	-	52
0.625	-	38

However, it's worth to notice, that number of edges in the ε=0.25 case was around 10 times higher than for ε=0.2, which seems to be perfect for these data – including modified procedure, where above ε=0.625 number of edges was growing fast.

The best clusterization is shown as a graph at image 2 (it doesn't contain the noise).

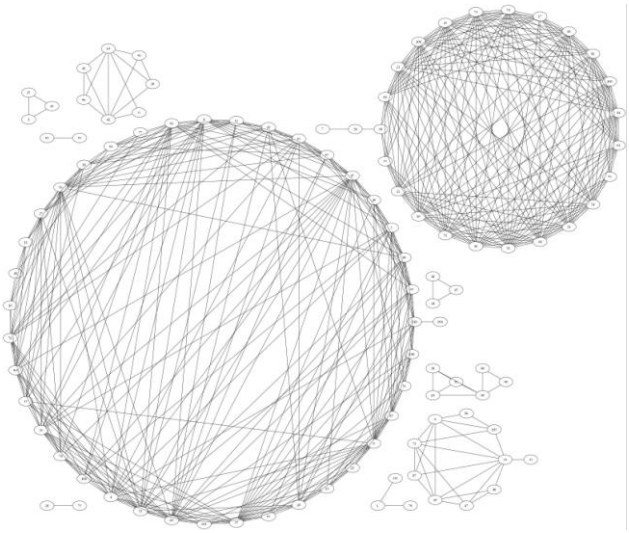

Image 2. Case with prices, standard DBSCAN procedure, ε=0.2

There has been chosen 28 offers from 113, which are the best from given set. Due to length of the data they cannot be shown in this work – they are available in [5].

5.2. CASE WITHOUT PRICES

Table 2 presents the results for the clusterization in the case without prices.

Table 2. Results of clusterization without prices

ε	Number of unclassified offers standard DBSCAN procedure	Number of unclassified offers modified DBSCAN procedure
0.1	61	-
0.125	**44**	-
0.15	38	-
0.4	-	44
0.45	-	36
0.5	-	33

For ε=0.125 there are two main clusters. This seems to be perfect parameter for this approach – higher ε links classes, lower ε produces many small classes. Probably for the higher amount of the data, smaller ε would be better. Best clusterization is shown as graph at image 3 (it doesn't contain the noise). It is impossible to give the number of offers to consider. Again, the accurate data are available in project report [5].

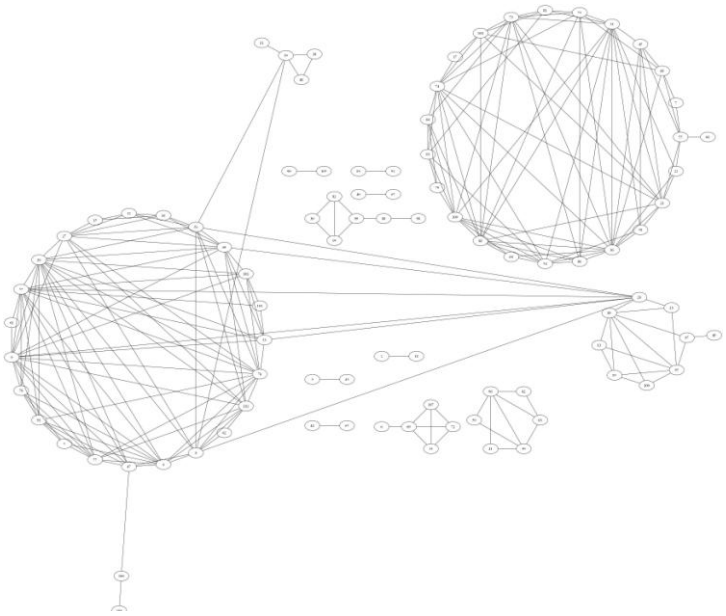

Image 2. Case without prices, standard DBSCAN procedure, ε=0.125

What is needed to add here is that from first three offers which can be extracted using this method, two of them were pointed out as well in the case with prices.

6. SUMMARY

The aim of this work was to show that it's possible to select the cost-effective offers using only pure mathematic clusterization methods to the set of real offers.

Though for the reason of the application which was generating graphs, the set of offers was bounded to 113, the extraction of the best offers was finished with full success. What's more, different approaches to the problem resulted in similar set of offers, what proves that all of them were correct.

After the computation, the same process has been performed for the set of 69 records. In that case, none of approaches was able to generate more than two clusters with satisfying level of noise. It's possible to infer that having big enough set of offers, and taking smaller ε, get better results will be obtained (as number of classes and their diversity), what causes big influence while considering the case 'without prices'. As well, in that case proposed measure should be more useful.

REFERENCES

[1] BERKHIN P., *Survey of Clustering Data Mining Techniques*, Accrue Software Inc. http://www.ee.ucr.edu/~barth/EE242/clustering_survey.pdf, 10 marca 2009

[2] ESTER M., KRIEGEL H., SANDER J., XU X., *A Density-Based Algorithm for Discovering Clustersin Large Spatial Databases with Noise*, http://www.cs.ualberta.ca/~joerg/papers/, 9 marca 2009

[3] LESSNAU A., *Klasteryzacja*, http://www.fizyka.umk.pl/~duch/zajecia/05SemMagInf/, 4 marca 2009

[4] NGUYEN H.S., *Clustering. Efektywne metody grupowania danych*, http://www.mimuw.edu.pl/~son/datamining/materials/w9_cluster.pdf, 8 marca 2009

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

[6] Dom i rynek – niezależne statystyki rynku nieruchomości, http://www.domirynek.pl/chart/time_avgprice/flat, 14 marca 2009

WYKORZYSTANIE METODY KLASTERYZACJI 'DBSCAN' DO ZBIORU OFERT SPRZEDAŻY MIESZKAŃ

W ostatnim okresie ceny nieruchomości w Polsce znacznie wzrosły. W związku z tym zakup mieszkania jest częściej traktowany jako inwestycja. Powoduje to zwiększenie zainteresowania matematycznymi, automatycznymi metodami wyznaczania opłacalności zakupu.

Niniejsza praca prezentuje wykorzystanie metody klasteryzacji DBSCAN do zbioru ofert sprzedaży mieszkań, która ma na celu wygenerowanie listy najlepszych ofert z początkowego zbioru.

W pracy został przedstawiony proces przygotowania danych, wyznaczenia cech oraz implementacja algorytmu. Do standardowej procedury DBSCAN zastosowano autorską metrykę. Została też przeprowadzona klasteryzacja bazująca na losowo wybranych 113 ofertach pochodzących z aglomeracji wrocławskiej. Otrzymane wyniki zaprezentowanej metody w porównaniu do panujących warunków na rynku nieruchomości okazały się być faktycznie lepsze.