PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1169 2007
TAKSONOMIA 14
Klasyfikacja i analiza danych – teoria i zastosowania

Agnieszka Mazur
Politechnika Łódzka
Dorota Witkowska
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie

ZASTOSOWANIE DRZEW KLASYFIKACYJNYCH W ANALIZIE RYNKU NIERUCHOMOŚCI*

1. Wstęp

Obserwowany wzrost zainteresowania rynkiem nieruchomości dotyczy szczególnie jednego z jego segmentów – rynku mieszkaniowego. Jak wynika z badań Głównego Urzędu Statystycznego, w roku 2005 średnia cena metra kwadratowego powierzchni mieszkania w skali całego kraju wzrosła o ok. 20% w stosunku do roku poprzedniego, natomiast liczba transakcji na rynku mieszkaniowym spadła w tym okresie o 13,23%[1].

Ocena zmian zachodzących na rynku nieruchomości wymaga szybkiego zdobywania i przetwarzania informacji niezbędnych do podjęcia optymalnych decyzji przez uczestników rynku, tj. sprzedających, kupujących, pośredników itp. Przydatne mogą się tutaj okazać metody statystycznej analizy wielowymiarowej, do których zalicza się drzewa klasyfikacyjne.

Celem badań jest zastosowanie drzew klasyfikacyjnych do grupowania mieszkań do wcześniej zdefiniowanych grup cenowych. Zmienną zależną w analizowanych modelach są ceny transakcyjne mieszkań, zmiennymi objaśniającymi (predyktorami) są zaś atrybuty tych mieszkań. Ocenę jakości klasyfikacji przeprowadzono na podstawie błędów klasyfikacji.

* Praca naukowa finansowana ze środków na naukę w latach 2005-2007 jako projekt badawczy.

[1] Notatka z dnia 6.07.2006, GUS.: http://www.stat.gov.pl/dane_spol-gosp/ceny_handel_uslugi/transakcje/2005/ZestawienieD2005.xls.

2. Opis obiektów

Badania empiryczne przeprowadzono, opierając się na rzeczywistych danych pochodzących z łódzkiego wtórnego rynku mieszkaniowego. Dane te zostały udostępnione przez agencję nieruchomości i dotyczą transakcji przeprowadzonych za pośrednictwem tej agencji w latach 2000-2001[2]. Wszystkie mieszkania opisane są przez następujące zmienne: x_1 – cena transakcyjna (w zł), x_2 – powierzchnia mieszkania (w m^2), x_3 – numer piętra, na którym znajduje się mieszkanie, x_4 – liczba pokoi, x_5 – rok budowy, x_6 – posiadanie telefonu (1 – tak, 0 – nie), x_7 – lokalizacja mieszkania, x_8 – rozkład mieszkania (1 – rozkładowe, 0 – amfilada), x_9 – występowanie balkonu (1 – jest, 0 – nie ma), x_{10} – rozmieszczenie mieszkania w budynku (1 – mieszkanie środkowe, 0 – mieszkanie szczytowe), x_{11} – ocena stanu technicznego (4 – bardzo dobry; 3 – dobry; 2 – do odświeżenia; 1 – średni; 0 – do remontu). Po wstępnej analizie danych wszystkie mieszkania zostały podzielone na cztery arbitralnie ustalone grupy cenowe (tab. 1).

Tabela 1. Struktura mieszkań według arbitralnie ustalonych grup cenowych

Cena w tys. zł	do 50	<50-75)	<75-100)	100 i powyżej
Grupa	I	II	III	IV
Liczebność	30	57	42	22

Źródło: opracowanie własne.

3. Drzewa klasyfikacyjne

Drzewa klasyfikacyjne dokonują podziału wielowymiarowej przestrzeni zmiennych $\mathbf{X}^m$, w której znajdują się klasyfikowane obiekty ($\mathbf{x}_1$, $\mathbf{x}_2$,..,$\mathbf{x}_n$), w taki sposób, aby uzyskać rozłączne segmenty tej przestrzeni. W tych segmentach znajdują się obiekty należące do tej samej klasy reprezentowanej przez zmienną zależną y. Procedura podziału ma charakter rekurencyjny i składa się z pięciu kroków [Gatnar, Walesiak 2004, s.108; Gatnar 2001, s. 24; Gatnar 1998, s. 169-170].

1. Sprawdzenie, czy przestrzeń $\mathbf{X}^m$, w której znajduje się zbiór uczący S, jest jednorodna pod względem wartości zmiennej zależnej y. Jeżeli tak, to należy zakończyć postępowanie[3], jeżeli nie, należy przejść do kroku 2.
2. Rozpatrzenie wszystkich możliwych sposobów podziału przestrzeni $\mathbf{X}^m$ na rozłączne segmenty R_1, R_2,..., R_k, uwzględniając wartości kolejno wybieranych zmiennych objaśniających.
3. Ocena jakości każdego z tych podziałów zgodnie z przyjętym kryterium homogeniczności oraz wybór najlepszego z nich.

[2] Średnia cena za m^2 mieszkania wynosi dla tego zbioru 1526 zł, według danych GUS, w I półroczu 2000 r. w województwie łódzkim wynosiła ona ok. 1489 zł, w roku 2004 – 773 zł, a w 2005 r. – 1178 zł.

[3] Procedurę można też zakończyć, gdy jest spełniony inny warunek stopu, np. minimalna liczba obiektów w powstałych segmentach, zob.: [Gatnar 1998, s. 176-180].

4. Podzielenie przestrzeni $\mathbf{X}^m$ wybranym sposobem.
5. Powtórzenie kroków 1-4 rekurencyjnie dla każdego z segmentów $R_1, R_2, ..., R_k$.

Optymalny model drzewa klasyfikacyjnego to taki, który daje mały błąd klasyfikacji i jednocześnie charakteryzuje się niewielką złożonością. W celu uzyskania takiego drzewa stosuje się takie metody, jak np. przycinanie krawędzi czy skracanie długości krawędzi, a następnie ocenia się błędy klasyfikacji dla zbioru uczącego (koszty resubstytucji) oraz zbioru testowego (koszty sprawdzianu krzyżowego).

W programach komputerowych do budowy drzew klasyfikacyjnych najczęściej wykorzystuje się algorytmy *CART* (*classification and regression trees*) i *QUEST* (*quick unbiased efficient statistical tree*)[4]. *CART* [Gatnar 1998, s. 182] jest algorytmem drzewa klasyfikacyjnego, polegającym na rekurencyjnym przeszukiwaniu wszystkich możliwych podziałów jednowymiarowych. W przypadku dużej liczby zmiennych predykcyjnych z wieloma poziomami działanie tego algorytmu może być czasochłonne, a także obciążone w kierunku wyboru zmiennych predykcyjnych, które mają więcej poziomów do podziałów. Dlatego często otrzymuje się najlepsze podziały z punktu widzenia najlepszych wyników klasyfikacji w zbiorze uczącym, jednak niekoniecznie w próbach sprawdzianu krzyżowego[5]. Algorytm *QUEST*, zamiast funkcji liniowych, wykorzystuje kwadratowe funkcje dyskryminacyjne do wyczerpującego poszukiwania podziałów jednowymiarowych [Gatnar, Walesiak 2004, s. 106]. Algorytm ten jest szybki i nieobciążony, jeżeli chodzi o wybór podziałów ze względu na liczbę poziomów predyktorów (predyktory z wieloma poziomami mają tendencję do generowania przypadkowych rozwiązań, które dobrze pasują do danych, ale mają słabą trafność predykcyjną).

4. Błędy klasyfikacji

Wyniki klasyfikacji można porównać, wyznaczając trzy rodzaje błędów klasyfikacji dla zbioru uczącego i testowego [Witkowska 2002, s. 92-93].

1. Ogólny błąd klasyfikacji:

$$E = \frac{n}{N} \cdot 100\%,$$

gdzie: n – liczba obiektów źle zaklasyfikowanych, N – liczba wszystkich obiektów.

2. Błąd klasyfikacji do innej klasy niż A_p (do której dany obiekt należy):

$$E_p = \frac{n_p}{N_p} \cdot 100\%,$$

gdzie: n_p – liczba obiektów, które faktycznie należą do klasy A_p, ale zostały błędnie zaklasyfikowane do innej klasy, N_p – liczba obiektów należących do klasy A_p.

[4] Algorytmy te są zaimplementowane m.in. w programie *Statistica*, który został wykorzystany do obliczeń.

[5] Podręcznik użytkownika programu *Statistica*.

3. Błąd klasyfikacji do klasy innej niż sąsiednia:

$$E_p^{\ s} = \frac{n_s}{N_p} \cdot 100\%,$$

gdzie: n_s – liczba obiektów, które faktycznie należą do klasy A_p, ale zostały błędnie zaklasyfikowane do klasy innej niż sąsiednia, N_p – liczba obiektów należących do klasy A_p.

5. Wyniki klasyfikacji

W badaniach wykorzystano zbiór 144 obiektów[6], który podzielony został losowo na zbiór uczący (105 przypadków) i zbiór testowy (39 obiektów). Za pomocą drzew klasyfikacyjnych dokonano grupowania mieszkań do wyróżnionych (tab. 1) czterech grup cenowych. W procesie budowy drzew klasyfikacyjnych zastosowano algorytmy *QUEST* i *CART*. W analizach przyjęto prawdopodobieństwa *a priori* proporcjonalne do liczebności grup cenowych oraz ustalono, że koszty błędnej klasyfikacji są równe we wszystkich grupach (tzn. błędne klasyfikacje oznaczają takie same straty we wszystkich klasach). Jako kryterium zatrzymania podziału zastosowano regułę przycinania przy błędzie złej klasyfikacji, co powoduje, że drzewo jest przycinane, gdy koszty klasyfikacji i złożoność drzewa są minimalne.

Tabela 2 zawiera statystyki drzew skonstruowanych za pomocą algorytmu *QUEST*, spośród których za najlepsze uznano drzewo o numerze 4. Składa się ono z czterech węzłów końcowych oraz charakteryzuje się najmniejszymi kosztami sprawdzianu krzyżowego, co oznacza, że dla drzew z większą liczbą węzłów końcowych błędy klasyfikacji w zbiorze testowym są już większe. Widać, że koszty resubstytucji (błędne klasyfikacje w próbie uczącej) maleją wraz ze wzrostem wielkości drzewa, natomiast koszty sprawdzianu krzyżowego na początku maleją wraz ze wzrostem wielkości drzewa, osiągają minimum i następnie zaczynają wzrastać, gdy wielkość drzewa staje się bardzo duża.

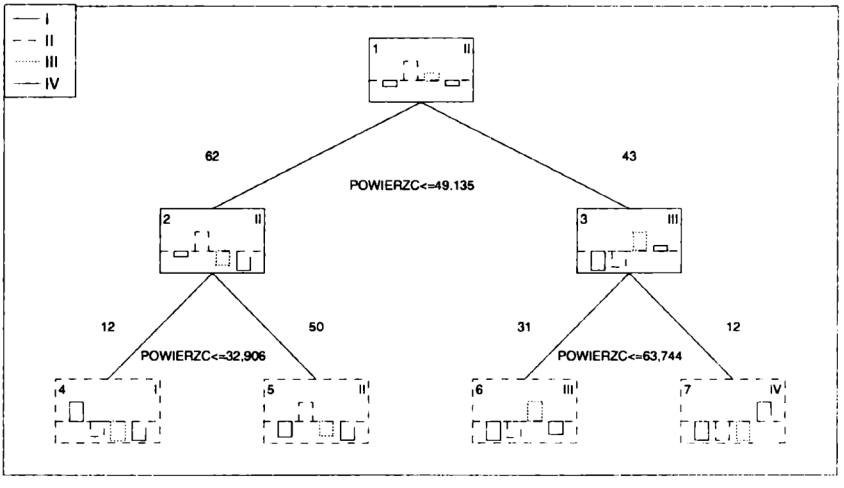
Tabela 2. Charakterystyki kolejnych drzew otrzymanych algorytmem *QUEST*

Nr	Liczba węzłów końcowych	Koszty sprawdzianu krzyżowego	Błąd standardowy kosztu sprawdzianu krzyżowego	Koszty resubstytucji	Złożoność drzewa
1	16	0,285714	0,044087	0,095238	0,000000
2	12	0,285714	0,044087	0,104762	0,002381
3	9	0,285714	0,044087	0,114286	0,003175
4	**4**	**0,200000**	**0,039036**	**0,152381**	**0,007619**
5	3	0,352381	0,046620	0,247619	0,095238
6	2	0,447619	0,048527	0,361905	0,114286
7	1	0,580952	0,048151	0,580952	0,219048

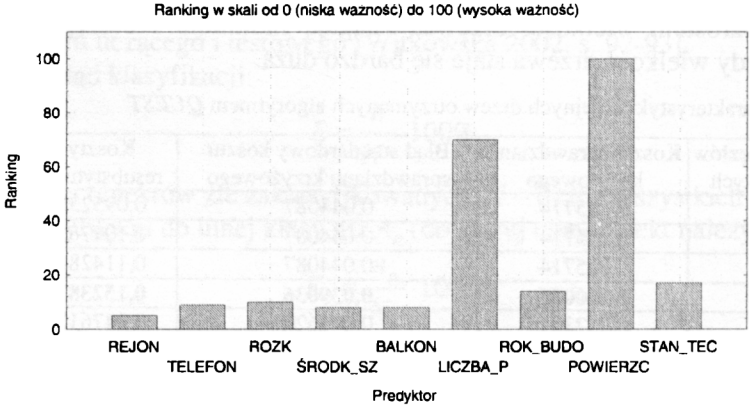
Źródło: obliczenia własne na podstawie *Statistica* v. 6.0.

[6] Z całego zbioru danych usunięto przypadki, w których występowały braki danych.

Na rysunku 1 przedstawiono wybrane (w tab. 2) drzewo klasyfikacyjne. Widać na nim, że początkowo cały zbiór obiektów został podzielony ze względu na powierzchnię mieszkań na dwa segmenty. W pierwszym z nich znalazły się 62 mieszkania o powierzchni do 49,135 m^2, w drugim – 43 mieszkania o powierzchni mniejszej. W następnym kroku każdy z tych segmentów został podzielony na kolejne dwa, również ze względu na tę samą cechę. W ten sposób powstały cztery grupy mieszkań o liczebnościach podanych w tab. 3. Można zauważyć, że drzewo dokonało klasyfikacji do grup cenowych na podstawie jednej zmiennej – powierzchni. Obiekty klasyfikowane do wyższych grup miały większą powierzchnię. Pozostałe atrybuty opisujące mieszkania okazały się mniej istotne; ranking ich ważności przedstawiono na rys. 2.



Rys. 1. Wybrane drzewo klasyfikacyjne
Źródło: opracowanie własne na podstawie *Statistica* v. 6.0.



Rys. 2. Ranking ważności predyktorów
Źródło: opracowanie własne na podstawie *Statistica* v. 6.0.

W tabeli 3 przedstawiono wyniki klasyfikacji. Dla zbioru uczącego uzyskano błąd klasyfikacji 15,24%, co oznacza, że 84,76% obiektów zostało poprawnie zaklasyfikowanych przez zbudowane drzewo klasyfikacyjne. Trochę większy błąd klasyfikacji uzyskano w odniesieniu do zbioru testowego, w którym odsetek poprawnych klasyfikacji wynosi 69,23%. Rozważając błędy klasyfikacji w poszczególnych grupach, można zauważyć, że najlepszy wynik uzyskano dla zbioru uczącego w grupie II (93,18% poprawnych klasyfikacji), a dla zbioru testowego – w grupie III (72,73% poprawnych klasyfikacji). Jak widać, wszystkie błędy klasyfikacji do innej niż sąsiednia grupy są równe zeru, co oznacza, że wszystkie błędnie zaklasyfikowane obiekty znalazły się w grupach sąsiednich.

Tabela 3. Wyniki klasyfikacji za pomocą drzewa klasyfikacyjnego metodą *QUEST*

Grupa cenowa	Zbiór uczący				Zbiór testowy			
	I	II	III	IV	I	II	III	IV
I	**11**	1	0	0	**10**	2	0	0
II	4	**41**	5	0	4	**7**	1	0
III	0	2	**25**	4	0	1	**8**	2
IV	0	0	0	**12**	0	0	2	**2**
Razem	15	44	30	16	14	10	11	4
Błędnie	4	3	5	4	4	3	3	2
100 - E_p (%)	73,33	93,18	83,33	75,00	71,43	70,00	72,73	50,00
100 - E (%)	84,76				69,23			

Źródło: obliczenia własne na podstawie *Statistica* v. 6.0.

Przy wykorzystaniu algorytmu *CART* (wybrane przez program) najlepsze drzewo miało podobne charakterystyki jak drzewo uzyskane algorytmem *QUEST*. W zbiorze uczącym poprawnie zaklasyfikowanych zostało 85,71% obiektów, a w zbiorze testowym – 61,54%. Najniższe błędy klasyfikacji uzyskano dla grupy II i III, trochę wyższe – w grupach I i IV. W zbiorze uczącym poprawność klasyfikacji w II grupie wynosi 90,91%, a w III – 93,33%, natomiast w zbiorze testowym – odpowiednio: 60 i 72,73%. W grupie I w zbiorze testowym 8 spośród 14 obiektów zostało poprawnie zaklasyfikowanych, a w grupie IV – tylko 2 spośród 4. Wszystkie błędnie zaklasyfikowane przypadki znalazły się w grupach sąsiednich. W tabeli 4 zestawiono błędy klasyfikacji dla obydwu algorytmów.

Tabela 4. Ocena jakości klasyfikacji za pomocą drzew klasyfikacyjnych

Algorytm		*QUEST*		*CART*	
Zbiór		uczący	testowy	uczący	testowy
100 - E (%)		84,76	69,23	85,71	61,54
100 - E_p (%)	I	73,33	71,43	66,67	57,14
	II	93,18	70,00	90,91	60,00
	III	83,33	72,73	93,33	72,73
	IV	75,00	50,00	75,00	50,00

Źródło: obliczenia własne.

Jak widać w tab. 4, w zbiorze uczącym nieznacznie lepszy wynik klasyfikacji otrzymano dla algorytmu *CART*, natomiast w zbiorze testowym więcej poprawnych klasyfikacji otrzymano dla algorytmu *QUEST*. W grupie III i IV w zbiorze testowym w obydwu przypadkach otrzymano takie same wyniki klasyfikacji. W grupie III poprawnie zaklasyfikowanych zostało 72,73% obiektów, w grupie IV zaś – tylko 50%.

W kolejnych eksperymentach wykorzystano różne zestawy zmiennych objaśniających (przykładowe wyniki przedstawiono w tab. 5), jednakże zawsze, gdy uwzględniano zmienną „powierzchnia", drzewa dawały wyniki porównywalne do podanych w tab. 4. Gdy w obliczeniach pominięto tę zmienną, a pozostawiono wszystkie pozostałe predykatory, uzyskano drzewo z dwoma węzłami końcowymi, które klasyfikowało mieszkania tylko do dwóch grup II i IV na podstawie zmiennej „liczba pokoi". W sytuacji, gdy ze zbioru predyktorów wyeliminowano obie najbardziej istotne zmienne, tj.: powierzchnię (x_2) i liczbę pokoi (x_4), drzewa klasyfikacyjne w ogóle nie rozwiązały problemu, wyodrębniając tylko II grupę.

Jak widać w tab. 5, poprawność klasyfikacji w modelach bez zmiennej x_2, opisującej powierzchnię, jest dużo gorsza. Spośród wszystkich drzew w tab. 5 najlepsze wyniki otrzymano dla drzewa z predyktorami: powierzchnia (x_2), rok budowy (x_5), ocena stanu technicznego (x_{11}), liczba pokoi (x_4) oraz występowanie balkonu (x_9). W zbiorze uczącym zaklasyfikowanych poprawnie zostało 84,76% obiektów, a w testowym – 67,27%. W przypadku pominięcia zmiennej opisującej powierzchnię najlepsze wyniki (tab. 5) otrzymano, gdy predyktorami były zmienne: x_3, x_4, x_5, x_6, x_8, x_9, x_{11} oraz $\{x_4, x_5, x_7\}$. Drzewo klasyfikacyjne dla pierwszego zestawu predyktorów składa się z 10 węzłów końcowych, drzewo uwzględniające zmienne: x_4, x_5, x_7 ma zaś cztery węzły końcowe.

Tabela 5. Ocena jakości klasyfikacji przy pomocy drzew klasyfikacyjnych dla różnych zestawów predyktorów

Zestaw predyktorów		$x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}$		$x_3, x_6, x_8, x_9, x_{10}, x_{11}$		$x_3, x_6, x_8, x_9, x_{10}$		$x_3, x_4, x_5, x_6, x_8, x_9, x_{11}$		$x_2, x_4, x_5, x_9, x_{11}$		x_4, x_5, x_7	
Zbiór		uczący	testowy	uczący	testowy	uczący	testowy	uczący	testowy	uczący	testowy	uczący	testowy
100 - E(%)		56,19	23,68	64,15	54,05	50,46	42,39	75,24	60,00	84,76	67,27	73,15	60,24
100 - E_p (%)	I	0	0	13,33	63,64	41,18	29,63	80,00	66,67	73,33	73,33	88,24	88,46
	II	100	88,89	70,45	59,26	45,45	37,84	88,64	50,00	93,18	68,75	81,82	50,00
	III	0	0	80,65	55,56	90,32	89,47	66,67	75,00	83,33	64,71	43,33	33,33
	IV	93,75	25	62,50	0	0	0	50,00	33,33	75,00	57,14	88,24	71,43

Źródło: obliczenia własne na podstawie *Statistica* v. 6.0.

Warto dodać, że przeprowadzono również klasyfikację obiektów, w której zmienną grupującą była cena metra kwadratowego mieszkania. Niestety otrzymane wyniki były dużo gorsze. Wszystkie obiekty były klasyfikowane tylko do dwóch grup lub do jednej grupy cen metra kwadratowego mieszkania.

6. Podsumowanie

Na podstawie otrzymanych wyników można stwierdzić, że drzewa klasyfikacyjne dały dość dobre wyniki klasyfikacji mieszkań do wyróżnionych czterech grup cenowych, gdy jednym z predyktorów była powierzchnia. W wyniku zastosowania dwóch różnych algorytmów (z uwzględnieniem wszystkich predyktorów) niewiele lepsze wyniki otrzymano dla algorytmu *QUEST*, który w zbiorze testowym zaklasyfikował poprawnie 69,23% mieszkań, podczas gdy algorytm *CART* – 61,54%. Drzewa klasyfikacyjne najsłabiej rozpoznały mieszkania z IV grupy cenowej zawierającej mieszkania najdroższe (tj. co najmniej za 100 tys. zł). Może to wynikać z faktu, iż w całym zbiorze danych tych mieszkań było najmniej. W pozostałych grupach drzewa klasyfikacyjne poprawnie rozpoznały w zbiorze testowym ok. 70-72% mieszkań. Należy również zauważyć, że wszystkie błędy klasyfikacji do grupy innej niż sąsiednia są równe zeru, co może być spowodowane tym, że obiekty w sąsiednich grupach są do siebie podobne.

W przeprowadzonych eksperymentach klasyfikacji mieszkań otrzymano drzewa o niezbyt złożonej strukturze, a cechą najbardziej istotną w procesie podziału okazała się powierzchnia mieszkania.

Literatura

Gatnar E. (2001), *Nieparametryczna metoda dyskryminacji i regresji*, PWN, Warszawa.

Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, PWN, Warszawa.

Gatnar E., Walesiak M. (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław.

Gospodarka mieszkaniowa w 2000 r. (2001), Informacje i Opracowania Statystyczne, GUS, Warszawa.

Transakcje kupna/sprzedaży nieruchomości w 2005 r., GUS, notatka informacyjna z dn. 6.07.2006 r.: http://www.stat.gov.pl/dane_spol-gosp/ceny_handel_uslugi/transakcje/2005/ZestawienieD2005.xls.

Witkowska D. (2002), *Sztuczne sieci neuronowe i metody statystyczne. Wybrane zagadnienia finansowe*, C.H. BECK, Warszawa.

APPLICATION OF CLASSIFICATION TREES TO THE REAL ESTATE MARKET ANALYSIS

Summary

In the research the data regarding the apartments, that were sold by their owners in Lodz region, are considered. Each apartment is described by the 11 attributes. In the paper we present the results of application the classification trees to classification of the apartments to four price groups.