PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Jacek Batóg
Uniwersytet Szczeciński

PRÓBA WYKORZYSTANIA PODEJŚCIA WIELOMODELOWEGO W KLASYFIKACJI JEDNOSTEK SAMORZĄDOWYCH

Streszczenie: W artykule podjęta została próba zweryfikowania hipotezy o uzyskiwaniu zadowalających efektów procesu agregacji modeli dyskryminacyjnych w przypadku bardzo małej liczebności prób uczących. Poszczególne modele bazowe budowane były przez rzutowanie na podprzestrzenie zmiennych objaśniających. W procesie agregacji zastosowana została architektura równoległa procesu agregacji wykorzystująca macierz wektorów prawdopodobieństw *a posteriori* i łączenie wyników predykcji za pomocą metody sumy. Rozważania teoretyczne zobrazowane zostały analizą porównawczą błędów predykcji uzyskiwanych w klasyfikacji jednostek samorządu terytorialnego.

1. Wstęp

Agregacja wyników uzyskiwanych w odrębnych badaniach, często również za pomocą zróżnicowanych metod, jest zjawiskiem stosunkowo znanym, zarówno w ekonomii, jak i w ramach innych nauk. Można tu wymienić na przykład zasadę maksymalizacji efektywności predykcji postulującą wykorzystywanie kilku alternatywnych metod prognozowania, wycenę nieruchomości za pomocą uśredniania wartości uzyskiwanych metodą dochodową, metodą księgowej wartości netto oraz metodą wartości odtworzeniowej, jak również stosunkowo popularną ostatnio koncepcję metaanalizy. Podstawowym celem agregacji rezultatów indywidualnych analiz jest zwiększenie prawdopodobieństwa podejmowania prawidłowych decyzji m.in. w zakresie przyporządkowania danego obiektu do określonej klasy lub wnioskowania o przyszłych wartościach badanych zjawisk.

Wielu autorów wskazuje, że koncepcja podejścia wielomodelowego oparta na agregacji modeli bazowych, stosowana w zagadnieniach klasyfikacyjnych, prowadzi do wzrostu dokładności predykcji w przypadku zbyt dobrze dopasowanych modeli do zbioru uczącego oraz stosowania różnych klas modeli bazowych, pozwala zwiększać liczbę modeli bazowych w przypadku małych zbiorów obserwacji, daje możliwość wykorzystania zróżnicowanych danych statystycznych oraz przez wykorzystanie tych danych uzyskiwanych w różnych okresach pozwala analizować ewolucję badanego zjawiska w czasie[1]. Przykładem potwierdzającym po-

[1] Zob. np. [Gatnar 2008].

wyższe stwierdzenia może być m.in. wykorzystanie ważonego podejścia wielomodelowego opartego na prawdopodobieństwach *a posteriori*[2] oraz zastosowanie metody baggingu w estymacji parametrów modeli regresji i drzew klasyfikacyjnych z wykorzystaniem algorytmu iteracyjnego *backfitting-like* w ramach procedury GAM-MM[3].

Należy jednak zauważyć, że zagregowany model drzew klasyfikacyjnych ma również pewne ograniczenia. Nie może on być m.in. przedstawiony w postaci drzewa klasyfikacyjnego, a więc nie jesteśmy w stanie uzyskać charakterystyki klas, która umożliwiałaby prostą interpretację uzyskiwanych wyników. Głównym zadaniem modelu tego typu jest predykcja przynależności nowych obiektów do danej klasy [Gatnar, Walesiak 2004, s. 124].

Dążąc do jak najwyższej trafności klasyfikacji, należy przestrzegać kilku podstawowych zasad. Agregacji powinny podlegać modele o niskiej dokładności, charakteryzujące się brakiem stabilności na zmianę rozkładu zmiennych objaśniających, a jednocześnie modele bazowe powinny być jak najbardziej zróżnicowane pod względem uzyskiwanych wyników predykcji. Duże znaczenie w podejściu wielomodelowym ma również wyznaczanie liczby modeli bazowych gwarantującej minimalizację błędu predykcji modelu zagregowanego. Przeprowadzane badania wskazują najczęściej, że optymalny jest w tym przypadku przedział od 50 do 200 modeli bazowych. Warto również wspomnieć, że zbyt duża liczba tych modeli powoduje pogarszanie się własności prognostycznych modelu końcowego[4].

2. Konstrukcja badania

W ramach przeprowadzonego badania weryfikacji poddana została hipoteza badawcza, według której wykorzystanie podejścia wielomodelowego pozwala zwiększyć dokładność klasyfikacji obiektów również w warunkach małej liczby modeli bazowych.

Klasyfikowanymi obiektami był zbiór 111 jednostek samorządu terytorialnego województwa zachodniopomorskiego, które były charakteryzowane przez cztery zróżnicowane trzyelementowe podzbiory zmiennych objaśniających opisujące stan finansów, poziom infrastruktury, rozwój demograficzny oraz strukturę podmiotów gospodarczych zachodniopomorskich gmin wiejskich, miejskich i miejsko-wiejskich w 2007 r. (zob. tab. 1). Typ jednostki samorządu terytorialnego stanowił w tym przypadku zmienną grupującą i wyznaczał 3 klasy klasyfikowanych obiektów[5].

W zastosowanym podejściu wielomodelowym poszczególne modele bazowe w postaci modeli dyskryminacyjnych uzyskiwane były przez rzutowanie na cztery

[2] Zob. [Stoica, Selén, Li 2004].
[3] *Generalised Additive Multi-Mixture Models*, zob. [Conversano 2002].
[4] Zob. np. [Gatnar 2008, s. 78-81].
[5] Ze względu na ich specyfikę nie uwzględniono miast na prawach powiatu.

podprzestrzenie zmiennych objaśniających[6] z wykorzystaniem równoległej architektury agregacji (zob. rys. 1). Proces agregacji wyników predykcji opierał się na macierzy wektorów prawdopodobieństw *a posteriori* (profilu decyzyjnego) i wykorzystaniu łączenia wyników predykcji za pomocą metody sumy (1) oraz dla porównania metody wartości maksymalnej (2):

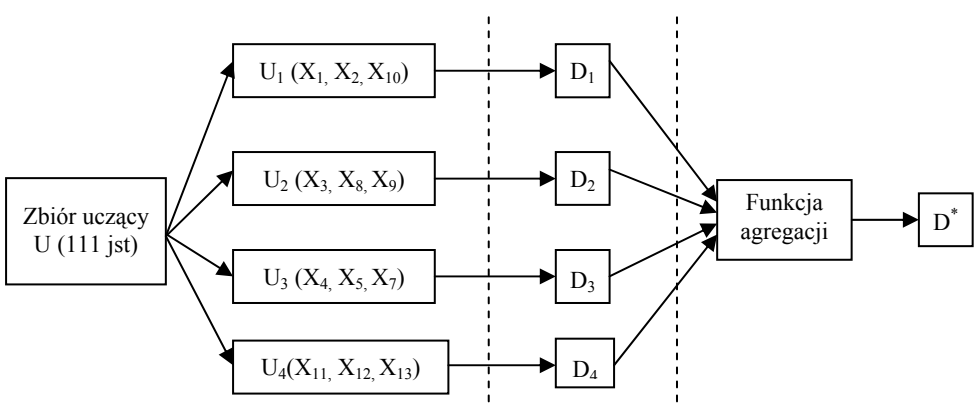
$$W_j(x_i) = \sum_{m=1}^{M} p_{m,j}(x_i)\,, \tag{1}$$

$$W_j(x_i) = \max_m \{p_{m,j}(x_i)\}\,. \tag{2}$$

Tabela 1. Wykorzystywane cechy diagnostyczne

Numer podzbioru	Nazwa zmiennej objaśniającej
1 (finansowy)	Dochody ogółem na 1 mieszkańca (X_1) Dochody własne na 1 mieszkańca (X_2) Dług publiczny jako % dochodu ogółem (X_{10})
2 (infrastrukturalny)	Odsetek ludności obsługiwanej przez oczyszczalnie (X_3) Odsetek ludności korzystającej z kanalizacji (X_8) Mieszkania nowo oddane do użytku na 1000 ludności (X_9)
3 (demograficzny)	Saldo migracji na 1000 ludności (X_4) Odsetek ludności w wieku produkcyjnym (X_5) Przyrost naturalny na 1000 ludności (X_7)
4 (gospodarczy)	Udział pracujących w usługach (X_{11}) Liczba firm na 1000 ludności (X_{12}) Liczba osób fizycznych prowadzących działalność gospodarczą na 1000 ludności (X_{13})

Źródło: opracowanie własne.



Rys. 1. Architektura stosowanego podejścia wielomodelowego

Źródło: opracowanie własne na podstawie [Gatnar 2008, s. 62].

[6] Szerzej na temat analizy dyskryminacyjnej traktuje m.in. praca [Jajuga 1990].

3. Wyniki empiryczne

Poniżej zaprezentowane zostały wyniki estymacji i trafność klasyfikacji wybranych modeli bazowych oraz modelu zagregowanego. W tabelach 2 i 3 przedstawiono rezultaty uzyskane dla trzeciego podzbioru zmiennych objaśniających (demograficznego), a w tab. 4 i 5 dla podzbioru czwartego (gospodarczego). Dla tego drugiego podzbioru cech diagnostycznych zobrazowano dodatkowo na rys. 2 rozrzut wartości kanonicznych.

Tabela 2. Wyniki estymacji modelu bazowego dla zmiennych demograficznych

Funkcja	R	λ Wilksa	χ^2	Stopnie swobody	p	Wartości własne	Wariancja (%)
1	0,427	0,809	22,69	6	0,0009	0,224	95,6
2	0,101	0,990	1,10	2	0,5780	0,010	4,4

Źródło: obliczenia własne.

Tabela 3. Macierz klasyfikacji dla modelu bazowego opartego na zmiennych demograficznych

	Trafność klasyfikacji (%)	M	W	MW
M	0,000	0	0	8
W	59,62	1	31	20
MW	70,59	0	15	36
Razem	60,36	1	46	64

Źródło: obliczenia własne.

Tabela 4. Wyniki estymacji modelu bazowego dla zmiennych gospodarczych

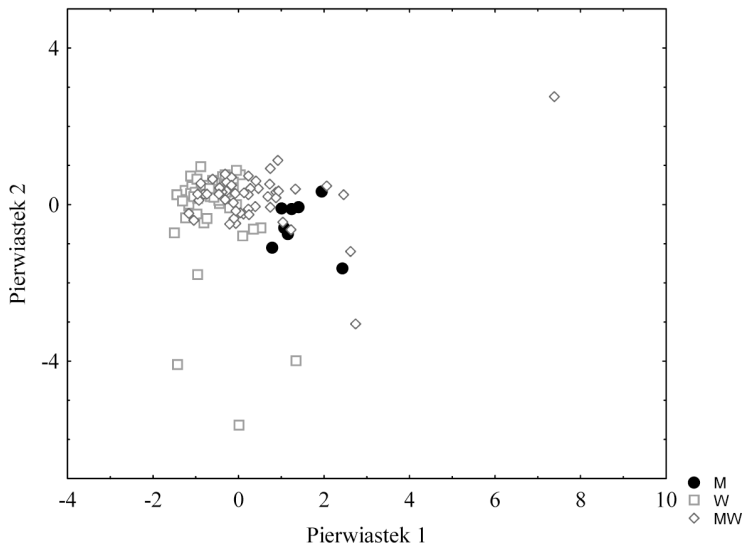
Funkcja	R	λ Wilksa	χ^2	Stopnie swobody	p	Wartości własne	Wariancja (%)
1	0,524	0,700	38,07	6	0,0000	0,379	91,5
2	0,185	0,966	3,71	2	0,1567	0,035	8,5

Źródło: obliczenia własne.

Tabela 5. Macierz klasyfikacji dla modelu bazowego opartego na zmiennych gospodarczych

	Trafność klasyfikacji (%)	M	W	MW
M	12,50	1	0	7
W	80,77	1	42	9
MW	52,94	3	21	27
Razem	63,06	5	63	43

Źródło: obliczenia własne.



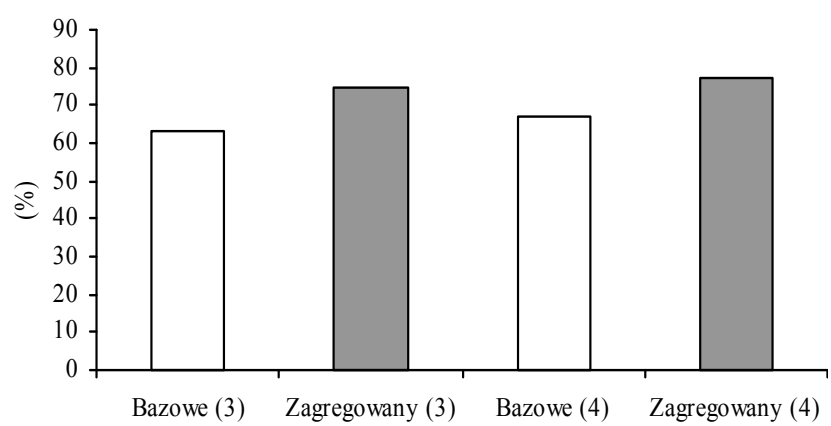
Rys. 2. Wykres rozrzutu wartości kanonicznych dla zmiennych X_{11}, X_{12} i X_{13}

Źródło: opracowanie własne.

Tabela 6. Macierz klasyfikacji dla modelu zagregowanego (metoda sumy)

	Trafność klasyfikacji (%)	M	W	MW
M	72,55	37	6	8
W	88,46	14	46	0
MW	0,00	8	0	0
Razem	74,77	51	52	8

Źródło: obliczenia własne.



Rys. 3. Porównanie trafności klasyfikacji modeli bazowych i zagregowanych

Źródło: opracowanie własne.

Wyniki określające ogólny błąd klasyfikacji w przypadku zmiennych finansowych i infrastrukturalnych kształtowały się odpowiednio na poziomie 61,26 oraz 68,47%.

Trafność klasyfikacji modelu zagregowanego uzyskanego z wykorzystaniem metody sumy (1) zamieszczono w tab. 6, a porównanie wyników klasyfikacji otrzymanych dla modeli bazowych oraz modelu zagregowanego przedstawiono na rys. 3, na którym uwzględniono również alternatywne modele bazowe zbudowane z wykorzystaniem podziału jednostek samorządu terytorialnego na 4 klasy na podstawie metody Warda.

W obydwu rozpatrywanych wariantach modeli bazowych wyraźnie widoczny jest wzrost trafności klasyfikacji dla modeli zagregowanych. Analizując powyższe rezultaty, warto również zwrócić uwagę na uzyskiwanie wyższej trafności w przypadku modeli, w których zmienna grupująca była tworzona za pomocą metody Warda, oraz przybliżone, lecz kształtujące się na nieznacznie niższym poziomie wyniki trafności klasyfikacji modeli zagregowanych przy wykorzystaniu metody wartości maksymalnej.

4. Podsumowanie

Wszystkie wyniki uzyskane w ramach przeprowadzonego badania pozwalają sformułować wniosek o znacznym spadku błędu ogólnego klasyfikacji jednostek samorządowych uzyskiwanego w przypadku budowy modeli zagregowanych. Otrzymane rezultaty były odporne zarówno na zmianę charakteru zmiennej grupującej (wykorzystane były dwa warianty tej zmiennej), jak i zróżnicowanie metod łączenia wyników predykcji (zastosowano metodę sumy oraz metodę wartości maksymalnej).

W celu potwierdzenia powyższych wniosków należałoby jednak przeprowadzić dalsze analizy, które pozwoliłyby ocenić przydatność podejścia wielomodelowego w przypadku agregacji małej liczby modeli bazowych przy stosowaniu architektur szeregowych i hybrydowych oraz różnicowania prawdopodobieństw *a priori*. Rozważać można również, czy jesteśmy w stanie minimalizować błąd klasyfikacji badanych obiektów przy wykorzystaniu próby testowej pozwalającej przypisać wagi poszczególnym modelom bazowym wyznaczone na podstawie obserwowanych błędów predykcji. Interpretacji poddane powinny zostać wyniki agregacji uzyskiwane na podstawie zróżnicowanych zbiorów danych.

Literatura

Conversano C., *Bagged mixtures of classifiers using model scoring criteria*, „Pattern Analysis and Applications" 2002 vol. 5.

Gatnar E., *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, PWN, Warszawa 2008.

Gatnar E., Walesiak M. (red.), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław 2004.
Jajuga K., *Statystyczna teoria rozpoznawania obrazów*, PWN, Warszawa 1990.
Stoica P., Selén Y., Li J., *Multi-model approach to model selection*, „Digital Signal Processing", vol. 14, Issue 5, September 2004.

THE APPLICATION OF A MULTI-MODEL APPROACH IN CLASSIFICATION OF ADMINISTRATIVE UNITS

Summary: In the paper, the author tried to assess whether a multi-model approach is effective in the case of small number of discriminant models. The individual classifiers were projection on different space of explanatory variables. The process of the aggregation was based on a parallel architecture of the overall system with the matrix of *a posteriori* probabilities and the sum method. The theoretical considerations were followed up by a comparative analysis of prediction errors in the process of classification of administrative units. In all considered cases a significant improvement of prediction accuracy of classification was observed.