Zeszyty Naukowe Uniwersytet Ekonomiczny w Krakowie

4(976)

ISSN 1898-6447
e-ISSN 2545-3238
Zesz. Nauk. UEK, 2018; 4(976): 123–143
https://doi.org/10.15678/ZNUEK.2018.0976.0408

Józef Pociecha

Współczesne zmiany narzędzi badań statystycznych

Streszczenie

W artykule zwrócono uwagę na współcześnie obserwowane zmiany narzędzi statystycznych służących badaniom naukowym w zakresie analizy i prognozowania procesów społeczno-ekonomicznych. Punktem wyjścia przeprowadzonych rozważań jest klasyczny schemat badań statystycznych w naukach ekonomicznych. Zwrócono uwagę na jego ograniczenia. Wskazano na współczesne metody analizy danych, oparte na regułach sztucznej inteligencji, które pomagają wyeliminować ograniczenie klasycznego schematu badań. Metody te należą do procedur uczenia nadzorowanego. Nawiązano do podstawowych metod klasyfikacji danych, jakimi są analiza dyskryminacyjna oraz model logitowy. Następnie scharakteryzowano te metody uczenia nadzorowanego, które również mogą mieć szersze zastosowanie w badaniach społeczno-ekonomicznych. Należą do nich: naiwny klasyfikator bayesowski, sieci bayesowskie, metoda *k*-najbliższych sąsiadów, metoda wektorów nośnych, klasyfikatory jądrowe, sztuczne sieci neuronowe, drzewa decyzyjne oraz podejście wielomodelowe (lasy losowe, *bagging*, *boosting*). Zwrócono uwagę, że i te metody podlegają jednak pewnym ograniczeniom.

Artykuł ma charakter przeglądowy i zawiera odniesienia do prac, w których zastosowano metody uczenia nadzorowanego w badaniach społeczno-ekonomicznych, opublikowanych w języku polskim.

Słowa kluczowe: klasyczny schemat badań statystycznych, metody klasyfikacji danych, uczenie nadzorowane, podejście wielomodelowe.
Klasyfikacja JEL: C10.

Józef Pociecha, Uniwersytet Ekonomiczny w Krakowie, Wydział Zarządzania, Katedra Statystyki, ul. Rakowicka 27, 31-510 Kraków, e-mail: pociecha@uek.krakow.pl

1. Klasyczny schemat badań statystycznych w naukach ekonomicznych

Według współczesnej klasyfikacji nauk nauki ekonomiczne są dziedziną w ramach obszaru nauk społecznych. Z naukami społecznymi dzielą metodologię badań społecznych. Metodologia ta zdominowana jest przez metody badań socjologicznych. Kompendium współczesnej wiedzy dotyczącej metodologii badań społecznych zawarte jest m.in. w pracy (Nowak 2007). Celem badań naukowych w naukach ekonomicznych jest lepsze zrozumienie zasad i praw kształtujących funkcjonowanie gospodarki zarówno w jej wymiarze mikroekonomicznym, jak i makroekonomicznym.

W badaniach naukowych stosuje się ścisłe reguły rozumowania zwane wnioskowaniem naukowym. Wnioskowanie to proces myślowy polegający na uznawaniu pewnych zdań za prawdziwe na podstawie innych, które zostały uznane wcześniej. Wyróżnia się dwa zasadnicze rodzaje wnioskowania: dedukcyjne oraz indukcyjne (patrz np. Kotarbiński 1990). Wnioskowanie dedukcyjne, nazywane popularnie – od ogółu do szczegółu, jest wnioskowaniem, w którym wniosek logicznie wynika z przesłanek. Wnioskowanie dedukcyjne przebiega według określonego prawa logicznego i dlatego jest niezawodne. Wnioskowanie indukcyjne, nazywane popularnie – od szczegółu do ogółu, to takie wnioskowanie, w którym na podstawie wielu przesłanek jednostkowych, stwierdzających, że poszczególne zbadane przedmioty pewnego rodzaju mają określoną cechę, dochodzi się, w sytuacji braku przesłanek negatywnych, do wniosku ogólnego, że każdy przedmiot tego rodzaju taką cechę posiada. Jeśli wiadomo, że nie ma innych przedmiotów danego rodzaju oprócz tych, które zostały wymienione w przesłankach jednostkowych, mówimy o wnioskowaniu przez indukcję zupełną, jeśli brak tej dodatkowej wiadomości – mówimy o wnioskowaniu przez indukcję niezupełną. Wnioskowanie przez indukcję zupełną jest wnioskowaniem niezawodnym. Wnioskowanie przez indukcję niezupełną nie gwarantuje prawdziwości wniosku, a jedynie go uprawdopodabnia.

W naukach społeczno-ekonomicznych stosowana jest zarówno metoda dedukcji, jak i indukcji. Punktem wyjścia w metodzie dedukcji jest postawienie hipotezy sformułowanej na podstawie rozumowania spekulatywnego. Następnie prowadzi się obserwację badanej rzeczywistości prowadzącą do przyjęcia lub odrzucenia wstępnie postawionej hipotezy. W końcowej fazie wnioskowania formułowana jest odpowiednia teoria. Metoda indukcji wychodzi od obserwacji badanej rzeczywistości. W następnej kolejności poszukiwany jest pewien wzorzec badanej rzeczywistości, co prowadzi do formułowania wstępnego wniosku w postaci hipotezy badawczej. Potwierdzenie sformułowanych hipotez może prowadzić do ich uogólnienia.

Specyficznym, a jednocześnie ważnym rodzajem wnioskowania indukcyjnego jest wnioskowanie statystyczne, zwane także indukcją statystyczną. Wnioskujemy statystycznie wówczas, gdy mając wiedzę na temat określonej cechy odnoszącej się do badanego przez nas zespołu obiektów (zwanego próbą), chcemy wyciągnąć wniosek dotyczący tej cechy w zespole obiektów znacznie większym niż badana próba, zwanym populacją generalną. Wyróżnia się dwie grupy metod uogólniania wyników, definiujące jednocześnie dwa działy wnioskowania statystycznego:

– estymacja statystyczna, polegająca na szacowaniu nieznanych rozkładów lub wartości ich parametrów w populacji,

– weryfikacja hipotez statystycznych, polegająca na sprawdzaniu poprawności przypuszczeń dotyczących postaci analitycznej rozkładu zmiennej lub parametrów rozkładu zmiennej w populacji, na podstawie wyników otrzymanych w próbie (patrz np. Hellwig 1998).

Metody wnioskowania statystycznego są przedmiotem gałęzi wiedzy matematycznej, która nazywana jest statystyką matematyczną.

Klasyczny schemat badań statystycznych w naukach ekonomicznych wychodzi od teorii ekonomicznej opisującej prawidłowości kształtowania się badanego procesu społeczno-gospodarczego. Następnie wskazania teorii ekonomii są formalizowane w postaci modelu matematycznego badanego procesu. W celu potwierdzenia lub zaprzeczenia istnienia prawidłowości wynikających z teorii ekonomii zbierane są odpowiednie dane statystyczne. Na ich podstawie dokonuje się estymacji parametrów modelu zbudowanego na podstawie wskazań teorii ekonomicznej. Końcowym etapem klasycznego schematu badań statystycznych jest ocena modelu pod względem stopnia jego dopasowania do badanej rzeczywistości. Alternatywnie zamiast estymować parametry teoretycznego modelu, możemy statystycznie weryfikować hipotezy o wartościach parametrów teoretycznego modelu w populacji lub weryfikować hipotezy nieparametryczne.

Przykładem takiego schematu badań statystycznych może być szacowanie parametrów mikroekonomicznej funkcji produkcji. Tutaj podstawą prowadzenia badań jest mikroekonomiczna teoria produkcji, opisująca relacje zachodzące pomiędzy nakładami czynników produkcji a ich efektem, czyli wytworzonym produktem. Kluczowym zagadnieniem jest wybór optymalnej kombinacji wyróżnionych czynników produkcji oraz skali produkcji. Podstawowym narzędziem analizy ekonomicznej w ramach teorii produkcji jest funkcja produkcji. Wyraża ona zależność między wielkością poniesionych nakładów (ilością czynników produkcji) na produkcję dóbr a osiągniętymi wynikami (ilość wytworzonego produktu). W najprostszej postaci, czyli przy założeniu dwóch czynników produkcji: pracy – L i kapitału – K, funkcja produkcji jest równa: $Y = f(L, K)$. Funkcja produkcji wyraża techniczno-bilansowy związek między strumieniami nakładów czynników produkcji a uzyskiwanym strumieniem produkcji (Y), przy

danej technologii produkcji. Spośród wielu postaci analitycznych funkcji produkcji jako podstawową przyjmuje się funkcję produkcji Cobba-Douglasa. Pierwszą polską monografią dotyczącą ekonometrycznej analizy procesu produkcyjnego jest praca (Pawłowski 1976).

Klasyczny schemat badań statystycznych w naukach ekonomicznych podlega jednak wielu ograniczeniom. Głównym źródłem ograniczeń jest różnorodność i konkurencyjność teorii ekonomicznych, które mogą być podstawą teoretyczną budowy i estymacji modelu ekonometrycznego. Przykładem może być spór pomiędzy neoklasyczną a keynesowską teorią ekonomii. Różna jest postać analityczna i częściowo inne czynniki są uwzględniane w modelach makroekonomicznych wychodzących z teorii neoklasycznej, a inna – w modelach bazujących na teorii Keynesa. Kolejnym ważnym ograniczeniem klasycznego schematu badań jest brak odpowiedniej teorii. Przykładowo, nie dysponujemy uznaną teorią ekonomiczną wpływu dopłat bezpośrednich na funkcjonowanie rolnictwa w krajach UE, nie możemy więc na tej podstawie budować modeli ekonometrycznych rozwoju rolnictwa w Polsce. Sposobem zaradzenia tym ograniczeniom jest wybór postaci analitycznej modelu metodą empiryczną – według kryterium najlepszego dopasowania do danych empirycznych. Konsekwencją takiego postępowania jest to, że każdy model dotyczy innej podstawy teoretycznej kształtowania się badanego zjawiska, a oszacowane modele mówią o kształtowaniu się badanego zjawiska w konkretnej rzeczywistości, bez możliwości ich bezpośredniego uogólniania. Tak zbudowane i oszacowane modele nie mówią o ogólnej teorii badanego zjawiska.

2. Analiza danych jako remedium na ograniczenia klasycznego schematu badań statystycznych

Analiza danych jest takim procesem ich przetwarzania, który ma na celu uzyskanie na ich podstawie informacji pozwalających na wyciąganie użytecznych wniosków. Głównym narzędziem analizy danych są metody statystyczne. Tradycyjnie rozumiana analiza danych obejmuje zastosowanie metod statystyki opisowej, wtedy wyciągane wnioski dotyczą analizowanego zbioru danych oraz statystyki matematycznej, analizowane są dane otrzymane z próby, a wyniki zostają uogólnione na całą populację metodami estymacji oraz weryfikacji hipotez statystycznych. Tradycyjne podejście zakłada arbitralny wybór metod statystycznych do analizy konkretnych zbiorów danych. Tradycyjna analiza danych mieści się w klasycznym schemacie badań statystycznych.

Współczesna analiza danych polega na zastosowaniu metod automatycznie poszukujących procedur pozwalających na przeprowadzenie optymalnej analizy danych. Należą one do obszaru wiedzy nazywanego uczeniem maszynowym.

Uczenie maszynowe (*machine learning*) jest analizą procesów uczenia się oraz tworzeniem systemów, które doskonalą swoje działanie na podstawie doświadczeń z przeszłości. Jest to dyscyplina naukowa wchodząca w skład nauk zajmujących się problematyką sztucznej inteligencji (Cichosz 2000).

Sztuczna inteligencja (*artificial intelligence*) jest dziedziną obejmującą wiedzę matematyczną, statystyczną, inżynierską i informatyczną. Zajmuje się ona przede wszystkim zagadnieniami matematycznego modelowania procesów przebiegających w organizmie człowieka. Tak zbudowane modele służą do opisu, a także prognozowania przebiegów procesów innych niż biologiczne, w tym procesów społeczno-gospodarczych. Drogą poszukiwania tego typu modeli są obliczenia i symulacje komputerowe. Z tego względu sztuczna inteligencja jest także rozumiana jako dział informatyki zajmujący się tworzeniem modeli zachowań inteligentnych oraz programów komputerowych symulujących te zachowania. Można ją też zdefiniować jako dział informatyki zajmujący się rozwiązywaniem problemów, które nie są efektywnie algorytmizowalne (Rutkowski 2009). W świetle powyższego sztuczną inteligencję można rozpatrywać w dwóch podstawowych znaczeniach: jako matematyczne modelowanie hipotetycznej inteligencji oraz jako technologię informatyczną służącą badaniom naukowym.

Głównym zadaniem badań nad sztuczną inteligencją jest konstruowanie maszyn i programów komputerowych zdolnych do realizacji wybranych funkcji umysłu i ludzkich zmysłów niepoddających się numerycznej algorytmizacji. Są to problemy informatyki na styku z neurologią, psychologią, kognitywistyką, systematyką oraz ze współczesną filozofią.

Rozróżniamy dwa zasadnicze typy uczenia maszynowego (Hastie, Tibshirani i Friedman 2009):

– uczenie nadzorowane (*supervised learning*) jest uczeniem maszynowym, które zakłada obecność ludzkiego nadzoru nad tworzeniem funkcji odwzorowującej wejście systemu na jego wyjście. Nadzór polega na podaniu programowi zbioru par wejście-wyjście (*input-output*) w celu nauczenia go podejmowania przyszłych decyzji. W tej sytuacji zakładamy udział człowieka w procesie uczenia;

– uczenie nienadzorowane (*unsupervised learning*) jest uczeniem maszynowym, które zakłada brak obecności dokładnego lub nawet przybliżonego wyjścia w danych uczących. Zadanie uczenia bez nadzoru polega na określeniu współzależności między cechami lub wykryciu wewnętrznej struktury zbioru danych. Przykładami uczenia nienadzorowanego są: analiza skupień (*cluster analysis*) czy analiza korespondencji. Metodami uczenia nienadzorowanego są metody taksonomiczne.

Dziedziną, w której z powodzeniem stosowane są metody uczenia maszynowego, są problemy statystycznej analizy i klasyfikacji danych. Z tego względu metody te alternatywnie nazywane są metodami uczenia statystycznego (Koronacki i Ćwik 2005, Krzyśko i in. 2008).

Główną funkcją nauki jest funkcja poznawcza. Tym samym naczelne cele i wartości realizowane w badaniach naukowych to przede wszystkim cele i wartości poznawcze. W obrębie nauk ścisłych i przyrodniczych napotykamy najczęściej przyczynowe wyjaśnienia zachodzenia określonych zjawisk. Nie jest to jednak jedyny rodzaj wyjaśniania, gdyż poza wyjaśnieniami przyczynowymi możliwe są także wyjaśnienia funkcjonalne, odwołujące się do pojęcia funkcji, oraz intencjonalne, odwołujące się do zamierzeń (intencji, zamierzonych celów) działających ludzi (Strawiński 2011). Metody uczenia statystycznego pozwalają realizować zasadniczą funkcję poznawczą, jaką jest klasyfikacja naukowa, czyli podział obiektów, przedmiotów, istot, osób czy zjawisk na jednostki klasyfikacyjne według określonych reguł i zasad. Podstawą klasyfikacji jest wprowadzenie jasnych i logicznych kryteriów podziału według typowych i unikalnych cech tego, co jest przedmiotem klasyfikacji. Celem klasyfikacji jest identyfikacja, czyli rozpoznanie określonego obiektu jako należącego do znanej nam klasy obiektów – o poznanych już wcześniej cechach, na podstawie których tę klasę wyróżniliśmy. Przejrzysta klasyfikacja pozwala przewidywać właściwości obiektów na podstawie ich pozycji w systemie. Dzięki takiemu podejściu uzyskujemy wiedzę o tym, co jest przedmiotem identyfikacji.

Klasyfikacja statystyczna jest rodzajem algorytmu statystycznego, który przydziela obserwacje statystyczne do klas, bazując na cechach tych obserwacji. Poklasyfikowanie badanych obiektów pozwala na poznanie ich wewnętrznej struktury, a tym samym wypełnia funkcję poznawczą nauki. Podstawowym narzędziem klasyfikacji obiektów w wielowymiarowej przestrzeni cech są metody taksonomiczne. Do podstawowych celów metod taksonomicznych należą: otrzymanie jednorodnych grup obiektów, ze względu na charakteryzujące je właściwości, co ułatwia ustalenie ich zasadniczych cech; redukcja dużej liczby informacji do kilku podstawowych kategorii, które mogą być traktowane jako przedmiot dalszej analizy, co pozwala na wyciągnięcie uogólniających wniosków; odkrycie nieznanej struktury analizowanych danych; zmniejszenie nakładów czasu i kosztów analiz poprzez ich ograniczenie do podstawowych zjawisk, procesów oraz kategorii. Metody taksonomiczne, pokazujące strukturę badanych obiektów, można podzielić na metody prowadzące do uporządkowania liniowego, polegającego na rzutowaniu obiektów na prostą w wielowymiarowej przestrzeni cech, oraz metody dające uporządkowanie nieliniowe, oparte na rzutowaniu tych punktów na płaszczyznę. Wśród metod taksonomicznych grupujących obiekty wyróżnia się metody grupowania bezpośredniego oraz metody grupowania iteracyjnego. Klasyczną monografią dotyczącą podstaw teoretycznych metod taksonomicznych oraz ich zastosowań w badaniach społeczno-ekonomicznych jest praca (Pociecha i in. 1988).

Uczenie nadzorowane obejmuje następujące etapy:

– zdefiniowanie zbioru danych, w postaci macierzy cech $\boldsymbol{X}$ opisujących jednostki statystyczne w pewnej zbiorowości statystycznej oraz wektora $\boldsymbol{y}$ będącego wartościami wynikowymi pewnej cechy kategorialnej lub metrycznej (*target variable*) odpowiadające obiektom $\boldsymbol{x}$,

– podział zbioru danych na zbiór uczący (*training set*) oraz zbiór testowy (*test set*) – problem proporcji podziału zbioru danych,

– określenie modelu uczenia pod nadzorem – wybór mechanizmu generującego, nauczyciela, klasyfikatora,

– badanie efektów klasyfikacji – analiza wrażliwości i specyficzności klasyfikacji.

Na każdym z wymienionych etapów uczenia nadzorowanego badacz podejmuje decyzje wpływające na rezultat prowadzonych badań. W tym sensie metody uczenia nadzorowanego nie są metodami automatycznie dającymi najlepsze rezultaty przeprowadzonych badań. Są jedynie metodami optymalizującymi wyniki w ramach przyjętych założeń dotyczących zbioru danych wejściowych oraz podjętych decyzji dotyczących wyboru mechanizmu generującego i wybranego klasyfikatora.

3. Metody uczenia nadzorowanego

3.1. Uwagi ogólne

Tradycyjnymi metodami klasyfikacji danych są: wielowymiarowa analiza dyskryminacyjna oraz model logitowy. Metody te pojawiły się, zanim sformułowane zostały procedury uczenia statystycznego oparte na zasadach sztucznej inteligencji. Jednakże zarówno analiza dyskryminacyjna, jak i model logitowy spełniają wymogi schematu uczenia nadzorowanego i jako pierwsze zostały wykorzystane w ramach procedur uczenia statystycznego.

Wielowymiarowa analiza dyskryminacyjna została zaproponowana przez R.A. Fishera (1936) jako narzędzie klasyfikacji indywidualnego obiektu do jednej z dwóch populacji, przez znalezienie takiej liniowej transformacji oryginalnych zmiennych, aby możliwie najlepiej odseparować od siebie obserwacje należące do różnych populacji. Prowadzi to do sformułowania zasad alokacji wyróżnionego obiektu do jednej z dwóch populacji. W klasycznym ujęciu funkcja dyskryminująca miała postać liniową, później próbowano stosować, na ogół bez zadowalających rezultatów, nieliniowe postacie funkcji dyskryminującej. Zasady klasyfikacji rozszerzano też na więcej niż dwie klasy. Metoda dyskryminacji liniowej Fishera stosowana była najpierw w badaniach przyrodniczych, następnie przeniesiona

została do badań psychologicznych oraz społeczno-ekonomicznych (Pociecha 2006).

Metoda wielowymiarowej analizy dyskryminacyjnej jest prezentowana w wielu pracach. Do klasycznych podręczników z tej dziedziny można zaliczyć (McLachlan 1992). Analiza dyskryminacyjna jest obecnie podstawową metodą klasyfikacji danych, wykorzystywaną do rozwiązywania wielu problemów decyzyjnych w analizie ekonomicznej i zarządzaniu. W literaturze można znaleźć bardzo wiele przykładów jej praktycznego zastosowania, nie zostaną one zatem tutaj przytoczone.

Zaletą analizy dyskryminacyjnej jest jej klarowna podstawa matematyczna oraz wysoka skuteczność klasyfikacji na homogenicznych zbiorach danych, zaś wadą jest fakt, że opiera się ona na stosunkowo restrykcyjnych założeniach i wykazuje słabą skuteczność na niehomogenicznych zbiorach danych.

Modele logitowe należą do klasycznych modeli klasyfikacji binarnej, to jest takich, w których zmienna objaśniana przyjmuje tylko dwie wartości. Model ten bada wpływ zmiennych objaśniających, które mogą być cechami jakościowymi lub ilościowymi, na zmienną objaśnianą o charakterze jakościowym, przy założeniu logistycznej postaci analitycznej natężenia poszczególnych zmiennych objaśniających. Jeśli zmienna objaśniana ma charakter dychotomiczny, to mamy do czynienia z modelem dwumianowym. Gdy zmienna objaśniana jest cechą jakościową wielowariantową, to mamy do czynienia z modelem wielomianowym uporządkowanym.

Z formalnego punktu widzenia model regresji logistycznej jest uogólnionym modelem liniowym (Hastie, Tibshirani i Friedman 2009), w którym wykorzystano formułę logitu jako funkcji wiążącej. Logitem nazywamy funkcję przekształcającą prawdopodobieństwo na logarytm ilorazu szans. Parametry modelu logitowego szacuje się metodą największej wiarygodności lub rzadziej uogólnioną metodą najmniejszych kwadratów.

Model logitowy jest obecnie jednym z najchętniej wykorzystywanych narzędzi statystycznych do analizy i prognozowania zjawisk gospodarczych. Może on być zarówno narzędziem klasycznej analizy statystycznej, jak również modelem uczonym w procesie nadzorowanym. W polskiej literaturze znajdujemy wiele przykładów jego zastosowania. Jako aktualny przykład zastosowania modelu logitowego w przewidywaniu bankructwa można wymienić pracę (Pawełek, Pociecha i Baryła 2016).

We współczesnych analizach społeczno-ekonomicznych mogą znaleźć szersze zastosowanie następujące metody uczenia nadzorowanego:

– naiwny klasyfikator bayesowski,
– sieci bayesowskie,
– metoda *k*-najbliższych sąsiadów,

– metoda wektorów nośnych,
– klasyfikatory jądrowe,
– sztuczne sieci neuronowe,
– drzewa decyzyjne,
– podejście wielomodelowe (lasy losowe, *bagging*, *boosting*).

W dalszej części niniejszej pracy zostaną one krótko scharakteryzowane, wraz ze wskazaniem przykładów ich zastosowań w badaniach społeczno-ekonomicznych.

3.2. Naiwny klasyfikator bayesowski

Naiwny klasyfikator bayesowski jest najprostszym klasyfikatorem probabilistycznym (Gatnar 2008). Funkcja klasyfikacyjna przyjmuje dwie wartości: 1 oraz –1. Naiwny klasyfikator bayesowski oparty jest na założeniu wzajemnej niezależności predyktorów, czyli zmiennych $\boldsymbol{X}$ opisujących pewną zbiorowość statystyczną. Cechy te mogą nie mieć związku z badaną rzeczywistością i dlatego są nazywane naiwnymi. Naiwny klasyfikator bayesowski oparty jest na modelu cech niezależnych, gdzie z prawdopodobieństwa *a priori* przynależności x do pewnej klasy wyprowadza się, korzystając z twierdzenia Bayesa, maksymalne prawdopodobieństwa *a posteriori*.

Naiwny klasyfikator bayesowski jest skutecznie uczony w trybie nadzorowanym. Pomimo naiwnego projektowania tego klasyfikatora oraz bardzo uproszczonych założeń, klasyfikator ten daje w wielu rzeczywistych sytuacjach wystarczająco dobre rezultaty. W praktyce naiwny klasyfikator bayesowski traktuje się jako punkt odniesienia w ocenie efektywności innych metod uczenia nadzorowanego.

Opis naiwnego klasyfikatora bayesowskiego można znaleźć m.in. w pracy (Koronacki i Ćwik 2005). Typowymi zastosowaniami naiwnego klasyfikatora bayesowskiego (Krzyśko i in. 2008) są przypadki, gdy kolumny macierzy danych $\boldsymbol{X}$ są zmiennymi dyskretnymi. W takim przypadku prawdopodobieństwa *a priori* estymowane są metodą częstościową z próby uczącej. Wykazano (Hastie, Tibshirani i Friedman 2009), że naiwny klasyfikator Bayesa może być użyteczny w zastosowaniach praktycznych, nawet gdy założenie o niezależności zmiennych nie jest spełnione. Klasyfikator ten jest powszechnie stosowany w zagadnieniach informatycznych oraz technicznych.

3.3. Sieci bayesowskie

Sieci bayesowskie, nazywane również bayesowskimi sieciami przekonań (*belief networks*), są graficznym modelem probabilistycznym, w którym strukturę stochastycznych zależności pomiędzy zmiennymi losowymi przedstawia się w postaci

grafu, w którym układ węzłów i krawędzi reprezentuje zmienne oraz zależności występujące pomiędzy nimi (Cichosz 2000).

W przypadku zmiennych nominalnych lub dyskretnych o stosunkowo niewielkiej liczbie możliwych wartości rozkłady warunkowe tych zmiennych przedstawia się w postaci tabel prawdopodobieństwa warunkowego i umieszcza się na schematach, zwyczajowo wewnątrz węzłów odpowiadających zmiennym. Znając zarówno strukturę grafu sieci, jak i odpowiednie rozkłady warunkowe oraz wykorzystując wzór Bayesa i twierdzenie o prawdopodobieństwie całkowitym, można obliczać najbardziej prawdopodobne konfiguracje stanów pewnych zmiennych sieci, przy znanych wartościach innych zmiennych.

Najtrudniejszym problemem dotyczącym praktycznego zastosowania sieci bayesowskich jest uczenie struktury sieci z danych empirycznych. Możliwe są tutaj trzy strategie. Pierwszą z nich jest uczenie struktury oparte na ograniczeniach, drugą stanowią algorytmy wartościujące, a trzecią – metody hybrydowe, łączące cechy dwóch poprzednich metod (Scutari 2010).

Bayesowskie sieci przekonań znajdują zastosowanie również w badaniach społeczno-ekonomicznych, jednak prace z tego zakresu są nieliczne. Do nich można zaliczyć artykuł (Gąska 2016).

3.4. Metoda *k*-najbliższych sąsiadów

Algorytm *k*-najbliższych sąsiadów (*k-nearest neighbours*) jest jednym z najbardziej znanych algorytmów regresji nieparametrycznej używanych od wielu lat w statystyce do prognozowania wartości zmiennych losowych. Jest on również często wykorzystywany dla celów klasyfikacji danych.

W procedurach uczenia maszynowego mamy dany zbiór uczący, zawierający obserwacje dotyczące wektora zmiennych objaśniających $\boldsymbol{X}$ oraz wartość zmiennej objaśnianej Y. Dana jest obserwacja C z przypisanym wektorem zmiennych objaśniających $\boldsymbol{X}$, dla której chcemy prognozować wartość zmiennej objaśnianej Y.

Algorytm *k*-najbliższych sąsiadów polega na:

– porównaniu wartości zmiennych objaśniających dla obserwacji C z wartościami tych zmiennych dla każdej obserwacji w zbiorze uczącym,

– wyborze k najbliższych do C obserwacji ze zbioru uczącego,

– uśrednieniu wartości zmiennej objaśnianej dla wybranych obserwacji, w wyniku czego uzyskujemy prognozę.

Definicja „najbliższych obserwacji" sprowadza się do minimalizacji pewnej metryki, mierzącej odległość pomiędzy wektorami zmiennych objaśniających dwóch obserwacji. Zwykle stosowana jest tu metryka euklidesowa lub metryka Mahalanobisa.

Z istoty metody *k*-najbliższych sąsiadów wynika, że w przypadku zbiorów uczących o dużej liczebności wyznaczenie odległości pomiędzy wszystkimi odległościami bardzo wydłuża czas poszukiwania najbliższych sąsiadów. W związku z tym konieczne jest ograniczanie liczby tych obserwacji do najważniejszych. W tym celu wykorzystywane są pewne algorytmy wstępnego przetwarzania zbioru uczącego. Inną niedogodnością metody *k*-średnich jest jej wrażliwość na tak zwane przekleństwo wielowymiarowości (Gatnar 2008).

Metoda najbliższych sąsiadów jako procedura nadzorowanego uczenia maszynowego opisana została w literaturze polskiej m.in. w pracach (Gatnar 2008) i (Krzyśko i in. 2008), zaś jako procedura hierarchicznego grupowania taksonomicznego pojawia się w wielu pracach z zakresu teorii i zastosowań taksonomii, wśród których można polecić np. (Pociecha i in. 1988). Przykładem zastosowania metody *k*-najbliższych sąsiadów w zagadnieniach ekonomicznych może być praca (Bartłomowicz 2010).

3.5. Metoda wektorów nośnych (SVM)

Metoda wektorów nośnych (*support vectors machines*) wywodzi się z prac F. Rosenblatta, który zaproponował algorytm poszukujący hiperpłaszczyzny, przy której sumaryczna odległość błędnie sklasyfikowanych obserwacji od granicy decyzyjnej byłaby możliwie najmniejsza. Metoda ta kompleksowo przedstawiona została w pracy (Vapnik 1995). W problemach liniowo separowalnych metoda SVM jednoznacznie określa położenie optymalnej hiperpłaszczyzny dyskryminującej. Umożliwia również znalezienie rozwiązania, gdy klas w próbie uczącej nie da się rozdzielić w sposób liniowy, czyli że można ją rozpatrywać w wariancie liniowym i nieliniowym. Hiperpłaszczyzna rozdzielająca klasy uzupełniona jest dwiema równoległymi prostymi wyznaczającymi pewien pas (margines) separujący klasy, nazywany wektorem nośnym.

Metoda SVM posiada wiele zalet. Do nich należy to, że stopień skomplikowania oraz pojemność metody jest niezależna od liczby uwzględnianych wymiarów. Bardzo dobra podbudowa statystyczno-teoretyczna metody opracowanej przez V. Vapnika pozwala na znajdowanie minimum globalnego, gdyż minimalizowana jest funkcja kwadratowa, co gwarantuje zawsze znalezienie minimum. Algorytm SVM jest bardzo wydajny i nie jest czuły na przetrenowanie. Pozwala też na uogólnianie otrzymywanych wyników. Poprzez wykorzystanie odpowiedniej funkcji jądrowej SVM wykazuje się dużą praktyczną skutecznością. Słabymi stronami metody SVM jest jej powolne uczenie się, szczególnie dokuczliwe przy dużej próbie uczącej. Przeważające zalety metody SVM przyczyniają się do jej dużej popularności w zastosowaniach praktycznych.

Metoda wektorów nośnych w sposób wyczerpujący przedstawiona została w monografii (James i in. 2013), a w języku polskim w pracach (Koronacki i Ćwik 2005) lub (Krzyśko i in. 2008). Zastosowania metody SVM w badaniach społeczno-ekonomicznych można natomiast znaleźć w publikacjach (Gąska 2013) czy (Trzęsiok 2010).

3.6. Klasyfikatory jądrowe

Klasyfikatory jądrowe (*kernel classifiers*), wykorzystujące funkcje jądrowe (Kulczycki 2005), stanowią ważną klasę funkcji potencjałowych. Są one metodą zbliżoną do metody SVM, gdyż jest to metoda uczenia nadzorowanego, jak również uwzględniają zasadę maksymalizacji marginesu oraz wykorzystują tak zwany *kernel trick* polegający na zastąpieniu iloczynów skalarnych funkcją jądrową (*kernel function*), a także dają możliwość jej interpretacji w kategoriach wnioskowania rozmytego. W takim przypadku metoda nazywana jest rozmytym klasyfikatorem maksymalnego marginesu (Gąska 2015). Podobnie jak w przypadku metody SVM, maksymalizuje ona margines oddzielający klasy przez odpowiednie dopasowanie parametrów, które są odpowiedzialne za nachylenie funkcji przynależności. Parametry te estymowane są metodą estymatorów jądrowych.

Estymator jądrowy gęstości jest rodzajem estymatora nieparametrycznego, przeznaczonym do wyznaczania gęstości rozkładu zmiennej losowej, na podstawie uzyskanej próby, czyli wartości, jakie badana zmienna przyjęła w trakcie dotychczasowych pomiarów. Estymator jądrowy gęstości jest często używany do analizy danych w licznych dziedzinach nauki i praktyki, zwłaszcza z zakresu technik informacyjnych, automatyki, zarządzania oraz wspomagania decyzji. Metody te często charakteryzują się wysoką skutecznością klasyfikacyjną wykazywaną w zagadnieniach praktycznych i w badaniach symulacyjnych.

Polska literatura dotycząca klasyfikatorów jądrowych jest dość obszerna, lecz związana z pracami w zakresie matematyki, informatyki oraz techniki. Jako podstawowe można uznać monografie (Kulczycki 2005) oraz (Krzyśko i in. 2008), zaś przykładem ich zastosowania w badaniach ekonomicznych jest praca (Gąska 2015).

3.7. Sztuczne sieci neuronowe

Koncepcja sztucznych sieci neuronowych powstała w latach 40. XX w., w wyniku poszukiwań matematycznego opisu działania komórek nerwowych (Tadeusiewicz 1993). Jest ona techniką informatyczną wzorowaną na strukturze i sposobie działania układów nerwowych organizmów żywych. Sztuczny neuron jest modelem swojego rzeczywistego odpowiednika. Jego zasadniczym celem jest przetworzenie informacji wejściowej, dostarczanej w postaci wektora o skończonej liczbie sygnałów wejściowych $x_1, \ldots, x_n$ w wartość wyjściową y. Przyjmuje się,

że zarówno wartości wejściowe neuronu, jak i wartość wyjściowa mają postać liczb rzeczywistych. Z każdym wejściem neuronu związany jest współczynnik nazywany wagą.

W celu stworzenia sieci neuronowej łączy się neurony w określony sposób. Zwykle neurony wchodzące w skład sieci tworzą warstwy, z których pierwsza nosi nazwę warstwy wejściowej, ostatnia – warstwy wyjściowej, zaś wszystkie warstwy znajdujące się pomiędzy nimi określane są jako warstwy ukryte. Wartości wejściowe sieci wprowadzane są na wejścia neuronów warstwy wejściowej. Następnie, poprzez istniejące połączenia, wartości wyjściowe neuronów jednej warstwy przekazywane są na wejścia elementów przetwarzających kolejnej warstwy. Wartości uzyskane na wyjściach neuronów ostatniej warstwy są wartościami wyjściowymi sieci.

Sposób funkcjonowania sieci neuronowej, gwarantujący prawidłowe rozwiązywanie postawionych przed nią problemów, uzależniony jest od dwóch podstawowych czynników:

– wartości współczynników wagowych neuronów składających się na sieć,

– struktury (topologii) sieci, która określana jest przez liczbę warstw, liczbę neuronów w poszczególnych warstwach, sposób połączeń neuronów oraz przyjęty model neuronu (sposób agregacji danych wejściowych, rodzaj zastosowanej funkcji aktywacji).

Ze względu na architekturę sieci neuronowych można wyróżnić jej trzy główne grupy (patrz np. Witkowska 2002). Pierwszą z nich stanowią sieci jednokierunkowe, a ich strukturę obrazuje acykliczny graf skierowany. Prostym przykładem sieci jednokierunkowej jest perceptron wielowarstwowy. Drugą grupę stanowią sieci rekurencyjne, w których dopuszcza się występowanie cykli. Typowym przykładem takiej sieci jest sieć Hopfielda. Trzecią grupę stanowią sieci komórkowe. Typowym przykładem tego typu sieci jest sieć Kohonena (SOM) (Kohonen 1995).

Polska literatura dotycząca sztucznych sieci neuronowych jest bardzo obszerna, począwszy od monografii (Tadeusiewicz 1993), i nie będzie tutaj przytaczana. Spośród wielu prac poświęconych zastosowaniom sieci neuronowych w badaniach społeczno-ekonomicznych można polecić monografie (Lula 1999) oraz (Migdał-Najman i Najman 2013).

3.8. Drzewa decyzyjne

Metoda drzew decyzyjnych została zaproponowana przez L. Breimana, J. Friedmana, R. Olshena i C Stone'a (1984). Polega ona na podziale przestrzeni obserwacji na rozłączne i dopełniające się wielowymiarowe kostki (segmenty) i dopasowaniu prostego modelu do każdej z nich oraz na wykorzystaniu przy tym danych z próby uczącej. W zależności od tego, do której kostki trafi nowa

obserwacja, decyzja co do wartości zmiennej wynikowej *y* podejmowana jest na podstawie jednego z lokalnie dopasowanych modeli. Drzewa decyzyjne są jednym z obrazów podziału rekurencyjnego badanego zbioru. Graficzną prezentacją metody podziału rekurencyjnego jest właśnie drzewo decyzyjne. Metoda drzew decyzyjnych jest zasadniczo graficzną metodą wspomagania procesu decyzyjnego stosowaną w teorii decyzji. Jest ona również stosowana w procesach uczenia nadzorowanego pozyskiwania wiedzy ze zbiorów danych.

Jeśli wyróżniona cecha *y* jest cechą nominalną, to reprezentujące ją drzewo nazywane jest drzewem klasyfikacyjnym, a jeśli jest to zmienna ciągła, to takie drzewo nazywamy drzewem regresyjnym. W zagadnieniach ekonomicznych częściej mamy do czynienia z drzewami klasyfikacyjnymi.

Istotnym problemem w zastosowaniach drzew regresyjnych lub klasyfikacyjnych jest wybór właściwego algorytmu konstrukcji drzewa. Spośród wielu znanych algorytmów szersze zastosowanie ma algorytm CLS (*conceptual learning systems*), będący podstawą wielu kolejnych algorytmów, oraz najczęściej stosowany algorytm CART (*classification and regression trees*). Praktyczne wykorzystanie tych algorytmów jest ułatwione dzięki temu, że znajdują się one w wielu pakietach statystycznych.

Modele drzew klasyfikacyjnych i regresyjnych posiadają wiele zalet. Najważniejszą zaletą jest łatwość interpretacji otrzymanego drzewa. Kolejną zaletą drzew jest możliwość klasyfikacji danych niepełnych. Ważna jest także możliwość ujęcia w modelach predyktorów mierzonych zarówno w skalach mocnych, jak i słabych, bez konieczności dokonywania ich transformacji. Wadą drzew decyzyjnych jest ich niestabilny charakter, objawiający się w skłonności do nadmiernego dopasowania do próby uczącej, co stało się impulsem do zaproponowania lasów losowych (Gatnar 2008).

Metoda drzew decyzyjnych jest dobrze opisana w polskiej literaturze. Przykładami mogą być monografie (Gatnar 2001) lub (Koronacki i Ćwik 2005). Wiele jest też prac dotyczących zastosowań drzew regresyjnych, a szczególnie klasyfikacyjnych w badaniach społeczno-ekonomicznych. Tutaj jako przykłady można przywołać monografie (Łapczyński 2010, Pociecha i in. 2014) lub artykuły (Chrzanowska i Drejerska 2015) lub (Witkowska 2015).

3.9. Podejście wielomodelowe

Wadą większości metod klasyfikacji danych z nauczycielem jest ich niestabilność, polegająca na skłonności do nadmiernego dopasowania do próby uczącej (przeuczenie sieci). Problem ten jest rozwiązywany drogą podejścia wielomodelowego. Istota podejścia wielomodelowego polega na łączeniu (agregacji) wielu klasyfikatorów bazowych w jeden model zagregowany, co poprawia skuteczność

klasyfikacyjną takiej konstrukcji. Model zagregowany z reguły jest bardziej dokładny niż którykolwiek z pojedynczych modeli wchodzących w jego skład, co pozwala na poprawę dokładności predykcji (Gatnar 2008).

Proces łączenia modeli przebiega w trzech fazach. Pierwsza polega na wyodrębnieniu ze zbioru uczącego *M* prób uczących. Próby te mogą zawierać podzbiory obserwacji wylosowane ze zbioru uczącego lub wszystkie obserwacje z tego zbioru, lecz scharakteryzowane przez różne (losowo wybrane) zmienne, co powoduje, że zbudowane na ich podstawie modele bazowe mogą się bardzo różnić wynikami predykcji. Faza druga polega na budowie *M* modeli bazowych. Modele te z reguły charakteryzują się małą dokładnością klasyfikacji oraz brakiem stabilności. W fazie trzeciej następuje łączenie wyników predykcji uzyskanych z modeli bazowych w model zagregowany (Gatnar 2008).

Istnieją dwa podstawowe warianty łączenia modeli bazowych w model zagregowany, co odpowiada dwóm rodzajom architektur agregacji:

1) architektura równoległa, gdy klasyfikatory bazowe są konstruowane niezależnie od siebie,

2) architektura szeregowa (sekwencyjna), gdy wyniki klasyfikacji jednej funkcji wpływają na konstrukcję kolejnych sekwencji klasyfikatorów bazowych.

Stosowana bywa także architektura hybrydowa, będąca połączeniem dwóch wymienionych architektur, oraz architektura warunkowa, która jest pewną odmianą architektury szeregowej (Gatnar 2008).

Spośród wielu znanych algorytmów podejścia wielomodelowego szersze zastosowanie mają: lasy losowe (*random forests*), metoda *bagging* oraz metoda *boosting*. Procedury te, jako mające wiele zastosowań w badaniach społeczno-ekonomicznych, zostaną syntetycznie scharakteryzowane w kolejnych częściach pracy.

Podstawową monografią w języku polskim przedstawiającą kompleksowo problemy podejścia wielomodelowego jest praca (Gatnar 2008). Aktualnymi przykładami zastosowania podejścia wielomodelowego w badaniach ekonomicznych są prace: (Dudek 2013) oraz (Pawełek i Grochowina 2017).

3.10. Lasy losowe

Metoda lasów losowych (*random forests*) zaproponowana przez L. Breimana (2001) dotyczy sytuacji, gdy jako modele bazowe wykorzystywane są drzewa klasyfikacyjne lub regresyjne. Obserwacje do *M* prób uczących losowane są bootstrapowo, to jest próby o liczności *n* losowane są zgodnie z rozkładem wyznaczonym przez dystrybuantę empiryczną oryginalnej próby i jednocześnie zmienne do modeli dobierane są losowo. Do budowy całego zbioru drzew, a następnie do ich klasyfikacji, wykorzystuje się zasadę głosowania wszystkich pośrednio

skonstruowanych klasyfikatorów. U podłoża omawianej metody leży zasada, że cecha „ważna" powinna mieć tę właściwość, iż losowa permutacja jej wartości w próbach prowadzi do istotnego pogorszenia skuteczności klasyfikacyjnej metody. Algorytm lasów losowych tworzy na zbiorze uczącym określoną przez użytkownika liczbę drzew klasyfikacyjnych. Aby zwiększyć przewagę lasu nad pojedynczym drzewem, pożądane jest zróżnicowanie (wariancja) drzew wchodzących w jego skład. Wariancja ta jest osiągana następującymi sposobami:

– losowy wybór zbioru uczącego dla każdego drzewa,

– losowy lub deterministyczny wybór zmiennych, na których dokonane zostaną cięcia (*splits*), zarówno dla całego drzewa, jak i pojedynczego cięcia,

– losowy wybór metody cięcia dla każdego drzewa,

– losowe zróżnicowanie minimalnego rozmiaru węzła (w zadanych granicach) dla każdego drzewa.

Stosowanie algorytmu drzew losowych umożliwia też tworzenie rankingu zmiennych pod względem ich ważności drogą ich losowych permutacji lub analizy wkładu danej cechy w obniżanie wartości współczynnika różnorodności węzłów, na etapie budowy poszczególnych drzew decyzyjnych.

Opis metody drzew losowych znajduje się w wielu pracach opublikowanych w języku polskim. Można tutaj przywołać m.in. monografię (Gatnar 2008) lub (Krzyśko i in. 2008). Przykładem zastosowania metody lasów losowych w zagadnieniach ekonomicznych jest np. praca (Fijorek i in. 2010). Lasy losowe częściej są stosowane w analizach medycznych. Zastosowania metody lasów losowych są ułatwione dzięki dostępnym programom w pakietach statystycznych.

3.11. Metoda bagging

Metoda agregacji bootstrapowej (metoda *bagging*) jest procedurą wielomodelową, krokową o architekturze typu równoległego zaproponowaną przez L. Breimana (1996). Od metody lasów losowych odróżnia ją to, że można ją zastosować wobec klasyfikatorów bazowych dowolnego typu, a nie tylko drzew. Metoda *bagging* polega na zbudowaniu M modeli bazowych na podstawie n-elementowych prób uczących wylosowanych ze zwracaniem, metodą bootstrapową, ze zbioru uczącego U. Obserwacje ze zbioru U, które nie znajdą się w żadnej z prób uczących, tworzą zbiór OOB (*out of bag*), wykorzystywany jako niezależny zbiór testowy.

Źródłem sukcesu metody agregacji bootstrapowej jest to, że redukuje ona wariancję modeli bazowych, które są „słabe", czyli charakteryzują się małą dokładnością oraz brakiem stabilności. Wykazuje się, że błąd modelu zagregowanego jest mniejszy od przeciętnego błędu modeli bazowych.

Algorytm metody *bagging* składa się z trzech kroków (Gatnar 2008):

1. Ustal liczbę modeli bazowych *M*.
2. Wykonaj dla każdego modelu bazowego następujące kroki:
   a) wylosuj próbę bootstrapową ze zbioru uczącego *U*, dla każdego modelu,
   b) zbuduj model bazowy na podstawie każdej z wylosowanych prób.
3. Dokonaj predykcji modelu zagregowanego dla obserwacji *x*, za pomocą modeli bazowych, stosując metodę głosowania, czyli przydzielania obserwacji *x* do tej klasy, którą wskazała największa liczba modeli.

Metoda *bagging* jest skuteczna w sytuacji niestabilności klasyfikatorów bazowych. Przez niestabilność klasyfikatorów należy rozumieć tendencję do otrzymywania różnych wyników klasyfikacji w przypadku wykorzystania różnych prób uczących. Jeśli procedura klasyfikacji prowadzi do powstawania niestabilnych klasyfikatorów, to metoda *bagging* przyczynia się do poprawy skuteczności klasyfikacyjnej. W przypadku klasyfikatorów stabilnych model zagregowany może być gorszy od najlepszych modeli bazowych (Rozmus 2013).

Podobnie jak opis metody drzew losowych, metoda *bagging* znajduje się w wielu pracach opublikowanych w języku polskim. Wyczerpującą jej prezentację można znaleźć w (Gatnar 2008) lub (Krzyśko i in. 2008). Natomiast przykładem zastosowania metody bagging w zagadnieniach ekonomicznych może być praca (Pełka 2015).

3.12. Metoda boosting

Metoda *boosting*, zaproponowana przez Y. Freunda i R. Schapirego (1997), oparta jest na architekturze szeregowej. Zawartość kolejnej próby uczącej zależy od predykcji poprzedzającego modelu, co oddaje nazwa *boosting* oznaczająca wzmacnianie czy poprawianie dokładności predykcji modelu zagregowanego w rezultacie kolejnych modyfikacji słabszych modeli bazowych. Wzmocnienie to uzyskuje się dzięki zastosowaniu podwójnego systemu wag. Pierwszy dotyczy obserwacji i polega na tym, że te obserwacje, które zostały błędnie sklasyfikowane przez model bazowy, uzyskują wyższe wagi. Drugi system wag dotyczy modeli bazowych i polega na przydzielaniu każdemu z modeli wag proporcjonalnych do jego błędu predykcji.

Podstawowym algorytmem metody *boosting* jest zaproponowany przez Y. Freunda i R. Schapirego (1997) algorytm *AdaBoost* (*adaptive boosting*). Występuje on w dwóch wersjach: z sekwencją losowego doboru próby (*resampling*) oraz z sekwencją ważenia obserwacji (*reweighting*).

Zaletą algorytmu *AdaBoost* jest po pierwsze możliwość redukcji obciążenia predykcji przez adaptacyjny system zmian wartości wag obserwacji, a po drugie

– zróżnicowanie modeli bazowych pomiędzy sobą, co pozwala zmniejszyć wariancję predykcji.

Metoda *boosting* w polskiej literaturze została wyczerpująco opisana m.in. w pracach: (Gatnar 2008) lub (Krzyśko i in. 2008). Znajduje ona zastosowanie przede wszystkim w zagadnieniach technicznych i medycznych, może być z powodzeniem stosowana także w analizach ekonomicznych, czego przykładem jest praca (Pełka 2012).

Przedstawione w powyższych punktach procedury uczenia statystycznego nie wyczerpują wszystkich możliwości praktycznego wykorzystania metod uczenia nadzorowanego, lecz są one jak dotychczas najczęściej stosowane w rozwiązywaniu problemów klasyfikacji obiektów natury społeczno-ekonomicznej.

4. Ograniczenia metod uczenia statystycznego

Wprowadzenie metod uczenia nadzorowanego do metodologii statystycznych badań procesów społeczno-ekonomicznych stanowi niewątpliwie postęp w stosunku do klasycznego schematu tych badań. Stało się to możliwe dzięki znacznemu wzrostowi mocy obliczeniowych dostępnych komputerów. Jednakże, wbrew potocznym oczekiwaniom, metody te nie rozwiązują automatycznie pojawiających się problemów badawczych.

Stosując metody uczenia statystycznego, należy mieć na uwadze, że one, tak jak inne metody, podlegają istotnym ograniczeniom. Po pierwsze, opierają się na założeniach mających na ogół ukryty charakter, charakterystycznych dla oferowanej procedury. Przykładem mogą być częściowo ukryte założenia dotyczące metod uczenia sieci neuronowych. W przypadku metod uczenia statystycznego możemy jedynie mówić, że podlegają one na ogół słabszym założeniom niż klasyczne metody badań statystycznych. Po drugie, podobnie jak zmienia się przedmiot badań społeczno-ekonomicznych, powinny zmieniać się przyjęte złożenia dotyczące jego analizy, co powoduje niestabilność przyjętych założeń badawczych w zmieniającej się rzeczywistości. Kolejnym problemem jest wybór właściwej procedury uczenia statystycznego. W literaturze znanych jest wiele procedur opartych na schemacie uczenia nadzorowanego, stąd pojawia się problem wyboru procedury analitycznej właściwej dla rozwiązania konkretnego problemu badawczego. Specyficznym problemem przy stosowaniu metod uczenia nadzorowanego jest ich niestabilność wynikająca z niebezpieczeństwa przeuczenia procedury, co powoduje nadmierne dopasowanie stosowanego modelu do danych empirycznych, niepozwalające jednak na ich wykorzystanie dla celów prognostycznych. Głównym powodem stosowania metod uczenia nadzorowanego w badaniach społeczno-ekonomicznych jest łatwość ich wykorzystania do celów predykcji.

W wyniku niestabilności przyjmowanych założeń badawczych oraz groźby przeuczenia procedur stopień ich predyktywności bywa jednak ograniczony.

Wymienione powyżej ograniczenia metod uczenia statystycznego składają się na główną słabość tych metod, jaką jest ograniczona możliwość uogólniania otrzymywanych tą drogą wyników, a tym samym formułowania na ich podstawie teorii ekonomicznych.

Metody analizy danych oparte na procedurach uczenia nadzorowanego należy umiejscowić pośrodku pomiędzy klasycznymi metodami analizy statystycznej a komputerowymi metodami obróbki danych. W równej mierze należą zarówno do statystyki, jak i informatyki. Żadna z tych dziedzin nie powinna ich zawłaszczać, gdyż są one typowym przykładem podejścia interdyscyplinarnego w badaniach naukowych.

Literatura

Bartłomowicz T. (2010), *Klasyfikacja nieruchomości metodą k-najbliższych sąsiadów*, „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu", nr 107, Taksonomia 17.

Breiman L. (1996), *Bagging Predictors*, „Machine Learning", vol. 24, nr 2, https://doi.org/10.1007/bf00058655.

Breiman L. (2001), *Random Forests*, „Machine Learning", vol. 45, nr 1, https://doi.org/10.1023/a:1010933404324.

Breiman L., Friedman J., Olshen R., Stone C. (1984), *Classification and Regression Trees*, CRC Press, London.

Chrzanowska M., Drejerska N. (2015), *Małe i średnie przedsiębiorstwa w strefie podmiejskiej Warszawy – określenie znaczenia lokalizacji z wykorzystaniem drzew klasyfikacyjnych*, „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu", nr 385, Taksonomia 25, https://doi.org/10.15611/pn.2015.385.05.

Cichosz P. (2000), *Systemy uczące się*, WNT, Warszawa.

Dudek A. (2013), *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

Fijorek K., Mróz K., Niedziela K., Fijorek D. (2010), *Prognozowanie cen energii elektrycznej na rynku dnia następnego metodami data mining*, „Rynek Energii", nr 12.

Fisher R.A. (1936), *The Use of Multiple Measurements in Taxonomic Problems*, „Annals of Eugenics", vol. 7, nr 2, https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Freund Y., Schapire R. (1997), *A Decision-theoretic Generalization of On-line Learning and an Application to Boosting*, „Journal of Computer and System Sciences", vol. 55, nr 1, https://doi.org/10.1006/jcss.1997.1504.

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

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

Gąska D. (2013), *Zastosowanie metody SVM do oceny ryzyka bankructwa i prognozowania upadłości przedsiębiorstw*, „Śląski Przegląd Statystyczny", nr 11.

Gąska D. (2015), *Prognozowanie bankructwa za pomocą klasyfikatorów rozmytych realizujących ideę maksymalnego marginesu*, „Śląski Przegląd Statystyczny", vol. 13, nr 19, https://doi.org/10.15611/sps.2015.13.06.
Gąska D. (2016), *Wykorzystanie sieci bayesowskich do prognozowania bankructwa firm*, „Śląski Przegląd Statystyczny", nr 14.
Hastie T., Tibshirani R., Friedman J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.
Hellwig Z. (1998), *Elementy rachunku prawdopodobieństwa i statystyki matematycznej*, PWN, Warszawa.
James G., Witten D., Hastie T., Tibshirani R. (2013), *An Introduction to Statistical Learning with Applications in R*, Springer, New York.
Kohonen T. (1995), *Sef-organizing Maps*, Springer, Berlin.
Koronacki J., Ćwik J. (2005), *Statystyczne systemy uczące się*, WNT, Warszawa.
Kotarbiński T. (1990), *Elementy teorii poznania, logiki formalnej i metodologii nauk,* Zakład Narodowy im. Ossolińskich, Wrocław.
Krzyśko M., Wołyński W., Górecki T., Skorzybut M. (2008), *Systemy uczące się – rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*, WNT, Warszawa.
Kulczycki P. (2005), *Estymatory jądrowe w analizie systemowej*, WNT, Warszawa.
Lula P. (1999), *Jednokierunkowe sieci neuronowe w modelowaniu zjawisk ekonomicznych*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.
Łapczyński M. (2010), *Drzewa klasyfikacyjne i regresyjne w badaniach marketingowych*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.
McLachlan G.J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.
Migdał-Najman K., Najman K. (2013), *Samouczące się sztuczne sieci neuronowe w grupowaniu i klasyfikacji danych. Teoria i zastosowania w ekonomii*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.
Nowak S. (2007), *Metodologia badań społecznych*, PWN, Warszawa.
Pawełek B., Grochowina D. (2017), *Podejście wielomodelowe w prognozowaniu zagrożenia przedsiębiorstw upadłością*, „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu", nr 468, Taksonomia 28.
Pawełek B., Pociecha J., Baryła M. (2016), *Dynamic Aspects of Bankruptcy Prediction. Logit Model for Manufacturing Firms in Poland* (w:) *Analysis of Large and Complex Data Studies in Classification*, red. A.F.X. Wilhelm, H.A. Kestler, Data Analysis and Knowledge Organization, Springer, Switzerland.
Pawłowski Z. (1976), *Ekonometryczna analiza procesu produkcyjnego*, PWN, Warszawa.
Pełka M. (2012), *Podejście wielomodelowe z wykorzystaniem metody boosting w analizie danych symbolicznych*, „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu", nr 242, Taksonomia 19.
Pełka M. (2015), *Adaptacja metody bagging z zastosowaniem klasyfikacji pojęciowej danych symbolicznych*, „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu", nr 384, Taksonomia 24, https://doi.org/10.15611/pn.2015.384.24.
Pociecha J. (2006), *Dyskryminacyjne metody klasyfikacji danych w prognozowaniu bankructwa firmy*, „Prace Naukowe Akademii Ekonomicznej we Wrocławiu", nr 1126, Taksonomia 13.

Pociecha J., Pawełek B., Baryła M., Augustyn S. (2014), *Statystyczne metody prognozowania bankructwa w zmieniającej się koniunkturze gospodarczej*, Fundacja Uniwersytetu Ekonomicznego w Krakowie, Kraków.
Pociecha J., Podolec B., Sokołowski A., Zając K. (1988), *Metody taksonomiczne w badaniach społeczno-ekonomicznych*, PWN, Warszawa.
Rozmus D. (2013), *Porównanie stabilności zagregowanych algorytmów taksonomicznych opartych na idei metody bagging*, „Studia Ekonomiczne", t. 133.
Rutkowski L. (2009), *Metody i techniki sztucznej inteligencji*, PWN, Warszawa.
Scutari M. (2010), *Learning Bayesian Networks with the bnlearn R Package*, „Journal of Statistical Software", vol. 35, nr 3, https://doi.org/10.18637/jss.v035.i03.
Strawiński W. (2011), *Funkcja i cele nauki – zarys problematyki metodologicznej*, „Zagadnienia Naukoznawstwa", vol. 3(189).
Tadeusiewicz R. (1993), *Sieci neuronowe*, Akademicka Oficyna Wydawnicza, Warszawa.
Trzęsiok M. (2010), *Wyodrębnianie reguł klasyfikacyjnych z modelu dyskryminacyjnego budowanego metodą wektorów nośnych*, „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu", nr 107, Taksonomia 27.
Vapnik V. (1995), *The Nature of Statistical Learning Theory*, Springer, Berlin.
Witkowska D. (2002), *Sztuczne sieci neuronowe i metody statystyczne. Wybrane zagadnienia finansowe*, C.H. Beck, Warszawa.
Witkowska D. (2015), *Wykorzystanie drzew klasyfikacyjnych do analizy zróżnicowania płac w Niemczech*, „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu", nr 384, Taksonomia 24, https://doi.org/10.15611/pn.2015.384.33.

Contemporary Changes in Statistical Research

(Abstract)

Many changes are observed in statistical tools for research in the field of analysis and the forecasting of socio-economic processes. The starting point of the considerations carried out is a classic scheme of statistical investigations in the economic sciences. Particular attention is paid to its limitations. Modern methods of data analysis, based on artificial intelligence, can help eliminate the limitations of the classical statistical investigations. These methods can be counted among supervised learning procedures. The paper next goes on to discuss the basic methods of data classification, including LDA and logit. Supervised learning methods that may have wider application in socio-economic studies are then presented. These include: the Naïve Bayes Classifier, Bayesian Networks, *k*-nearest neighbours, vector support machines, kernel classifiers, artificial neural networks, decision trees, and a multi-model approach (random forests, bagging, boosting). However, these methods are also subject to certain restrictions.

The article is an overview and contains references to works in which supervised learning methods have been applied in socio-economic studies.

Keywords: classical scheme of statistical investigations, data classification methods, supervised learning, multimodel approach.