Załącznik nr 3. Autoreferat

dr Marcin Pełka

Autoreferat przedstawiający opis dorobku i osiągnięć naukowych

Spis treści

1. Imię i nazwisko

Marcin Pełka

2. Posiadane dyplomy i stopnie naukowe

2002 r. – uzyskanie tytułu magistra ekonomii na Wydziale Gospodarki Regionalnej i Turystyki, kierunek: Ekonomia, specjalność: Bankowość i ubezpieczenia. Tytuł pracy magisterskiej: „Rynek polskich funduszy inwestycyjnych w latach 1998-2002".

2007 r. – uzyskanie stopnia doktora nauk ekonomicznych w zakresie ekonomii na Wydziale Gospodarki Regionalnej i Turystyki w Jeleniej Górze Akademii Ekonomicznej im. Oskara Langego we Wrocławiu. Temat rozprawy doktorskiej: „Analiza danych symbolicznych i jej wykorzystanie w badaniach marketingowych". Promotorem w przewodzie doktorskim był prof. zw. dr hab. Marek Walesiak, a recenzentami byli prof. zw. dr hab. Eugeniusz Gatnar i dr hab. Andrzej Bąk. Stopień doktora został mi nadany Uchwałą Rady Wydziału Gospodarki Regionalnej i Turystyki w Jeleniej Górze 29 września 1998 r.

3. informacje o dotychczasowym zatrudnieniu

Od października 2006 r. byłem zatrudniony jako asystent w Katedrze Ekonometrii i Informatyki na Wydziale Gospodarki Regionalnej i Turystyki Uniwersytetu Ekonomicznego we Wrocławiu.

Od października 2008 r. jestem zatrudniony jako adiunkt w Katedrze Ekonometrii i Informatyki na Wydziale Ekonomii i Finansów Uniwersytetu Ekonomicznego we Wrocławiu.

4. Omówienie osiągnięcia naukowego, o których mowa w art. 219 ust. 1 pkt. 2 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2021 r. poz. 478 z późn. zm.)

4.1. Tytuł osiągnięcia i skład cyklu publikacji powiązanych tematycznie

Jako osiągnięcie naukowe przedstawiam cykl dwudziestu dwóch tematycznie powiązanych publikacji pod zbiorczym tytułem:

Podejście wielomodelowe analizy danych symbolicznych w badaniach ekonomicznych

W skład cyklu publikacji wchodzą następujące pozycje:

[1] Pełka M., Dudek A. (2009), *Effectiveness of symbolic classification trees vs. noisy variables,* Acta Universitatis Lodziensis. Folia Oeconomica nr 228, s. 173-179.

[2] Pełka M. (2009), *Sieci neuronowe dla danych symbolicznych: perceptron wielowarstwowy*, Prace Naukowe UE we Wrocławiu nr 47, s. 214-222.

[3] Pełka M. (2010), *K-nearest neighbour classification for symbolic data*, Acta Universitatis Lodziensis. Folia Oeconomica, nr 235, s. 171-176.

[4] Pełka M. (2017), *Wielomodelowa klasyfikacja spektralna danych symbolicznych.* Pace Naukowe UE we Wrocławiu nr 468, s. 180-187.

[5] Pełka M. (2015), *An adaptation of COBWEB for symbolic data case*. Statistica, Vol. 75, No. 3, 265-273.

[6] Pełka M. (2011), *Podejście wielomodelowe w analizie danych symbolicznych – metoda bagging*, PN UE we Wrocławiu nr 176, s. 375-382.

[7] Pełka M. (2012), *Podejście wielomodelowe z wykorzystaniem metody boosting w analizie danych symbolicznych*, PN UE we Wrocławiu nr 242, s. 315-322.

[8] Pełka M. (2014), *Podejście wielomodelowe w regresji danych symbolicznych interwałowych*, PN UE we Wrocławiu. Ekonometria 4 (46), s. 211-220.

[9] Pełka M. (2012), *Ensemble approach for clustering of interval-valued symbolic data*, Statistics in Transition, Volume 13, Number 2, s. 335-342.

[10] Pełka M. (2013), *Podejście wielomodelowe analizy danych symbolicznych w ocenie pozycji produktów na rynku*, Ekonometria 2(40), s. 95-102.

[11] Pełka M. (2014a), *Symbolic cluster ensemble based on co-association matrix versus noisy variables and outliers*, [w:] Spiliopoulou M., Schmidt-Thieme L., Janning R. (Eds.), *Data analysis, machine learning and knowledge discovery*, Springer-Verlag, Berlin-Heidelberg, s. 209-216.

[12] Pełka M. (2014b), *Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym*. PN UE we Wrocławiu nr 327, s. 202-209.

[13] Pełka M. (2015), *Adaptacja metody bagging z zastosowaniem klasyfikacji pojęciowej danych symbolicznych*. PN UE we Wrocławiu nr 384, s. 227-235.

[14] Pełka M. (2016), *A comparison study for spectral, ensemble and spectral mean-shift clustering approaches for interval-valued symbolic data*. [W:] Wilhelm A., Kestler H. (red.), Analysis of Large and Complex Data, Springer-Verlag, Berlin-Heidelberg, s. 137-146.

[15] Pełka M. (2017), *Wielomodelowa klasyfikacja spektralna danych symbolicznych*. PN UE we Wrocławiu nr 468, s. 180-187.

[16] Pełka M. (2017), *Klasyfikacja wielomodelowa danych symbolicznych w badaniu innowacyjności krajów Unii Europejskiej*. PN UE we Wrocławiu Ekonometria 2 (56), s. 42-51.

[17] Pełka M. (2018), *Analysis of innovations in the European Union via ensemble symbolic density clustering*, Econometrics. Advances in Applied Data Analysis, vol. 22, no. 3, s. 84-98.

[18] Pełka M. (2019) *Assessment of the Development of the European OECD Countries with the Application of Linear Ordering and Ensemble Clustering of Symbolic Data*. Folia Oeconomica Stetinensia volume 19, issue 2, s. 117–133.

[19] Pełka M., Rybicka A. (2019), *Hybrid Conjoint Analysis – Symbolic Decision Tree Model for Customer Churn Prediction Model.* W: Vision 2025: Education Excellence and Management of Innovations through Sustainable Economic Competitive Advantage. Proceedings of the 34th International Business Information Management Association Conference (IBIMA) / Soliman Khalid S. (red.), 2019, International Business Information Management Association, s.12435-12441, ISBN 9780999855133

[20] Pełka M. (2019), *Symbolic decision stumps in individual credit scoring*. Bank i Kredyt 50(6), s. 512-528.

[21] Pełka M., Rybicka A. (2020), *Symbolic Ensemble Clustering And Linear Ordering Of European Countries According To Their Economic Freedom*, w: Education Excellence and Innovation Management: A 2025 Vision to Sustain Economic Development during Global Challenges / Soliman Khalid S. (red.), International Business Information Management Association (IBIMA), ISBN 9780999855141, ss. 4788-4797.

[22] Pełka M. (2020), *Improving Classification Accuracy of Ensemble Learning for Symbolic Data Trough Neural Networks' Feature Extraction*, [w:] K. Jajuga, J. Batóg, M. Walesiak (red.), Classification and data analysis. Theory and applications. Springer International, ss. 73-84.

4.2. Wprowadzenie

Pojęcie modelu i modelowania jest bardzo często używanym pojęciem. Mimo, że pojęcie modelu jest często stosowane, to jego definicja nie jest prosta. Model jest pewną konstrukcją teoretyczną, która podlega analizowaniu w miejsce rzeczywistego obiektu czy zjawiska. Pozwala on na lepsze zrozumienie charakteru tegoż obiektu czy zjawiska. Jest to w znacznym stopniu uproszczony obraz rozpatrywanego systemu ekonomicznego, społecznego czy fizycznego (Gatnar 1993, s. 11).

W literaturze przedmiotu odnaleźć wiele różnych klasyfikacji modeli (zob. np. Gatnar 1993, s. 14-36). Można odnaleźć także dyskusję nad zasadnością stosowania podejścia ilościowego, jakościowego i symbolicznego w ekonomii (Dudek 2013, s. 16-34).

Pojawia się zatem problem wyboru odpowiedniego modelu w sytuacji, gdy na podstawie samych tylko danych, ich własności, nie można jednoznacznie stwierdzić, który z modeli byłby tym najbardziej adekwatnym. Odpowiedzią może być w takim przypadku podejście wielomodelowe (por. m.in. Gatnar 2008; Zhi-Hua 2012; Kuncheva 2014; Polikar 2006; Rokach 2010).

Podejście wielomodelowe polega na zastosowaniu do danego problemu wielu różnorodnych modeli, a następnie na połączeniu ich wyników w jeden, zagregowany, model wynikowy. Model zagregowany jest także dużo dokładniejszy niż którykolwiek z modeli bazowych, które wchodzą w jego skład. Innymi przesłankami, poza problematyką wyboru odpowiedniego modelu, które przemawiają za stosowaniem podejścia wielomodelowego są z pewnością problemy związane z wielkością zbioru danych. W przypadku bardzo dużych zbiorów danych, których analizowanie za pomocą jednego modelu może być utrudnione, podejście wielomodelowe pozwala podzielić taki zbiór na mniejsze części jak ma to miejsce w metodzie *boosting* (Efron 1979). Natomiast w przypadku zbiorów danych o niewielkiej liczebności podejście wielomodelowe pozwala wykorzystywać wielokrotnie ten sam, niewielki, zbiór danych (Polikar 2006).

Wśród innych przesłanek przemawiających za podejściem wielomodelowym warto wskazać z pewnością możliwość podziału zbioru danych, którego obiekty tworzą złożoną, skomplikowaną, strukturę na mniejsze grupy, w których odkrycie istniejących zależności będzie o wiele łatwiejsze. Dodatkowo podejście wielomodelowe, czyli łączenie wyników wielu modeli, w jeden model zagregowany stosowane jest w wielu innych naukach – np. w medycynie, gdzie do postawienia diagnozy na temat choroby stosowanych jest wiele badań czy testów.

W ramach przesłanek filozoficznych przemawiających za stosowaniem podejścia wielomodelowego w literaturze przedmiotu wskazuje się zasadę konieczności gromadzenia i analizowania wielu obserwacji (wyjaśnień zjawiska), którą propagował Epikur [Asmis 1984]. Za podobną zasadę, lecz w innym brzmieniu można by nawet uznać naiwny falsyfikacjonizm Poppera (Vermaas 2014).

Właśnie możliwość zastosowania wielu metod i łączenia ich wyników w jeden model oraz przyjazność tego podejścia stanowiła dla mnie przesłankę do zainteresowania się tym podejściem na gruncie analizy danych symbolicznych.

W literaturze przedmiotu podejście wielomodelowe obejmuje zarówno zagadnienia wzorcowe (Gatnar 2008; Kuncheva 2014, Zhi-Hua 2012) jak i bezwzorcowe (Vega-Pons i Ruiz-Shulcloper 2011; Ghaemi i in. 2009). Podobnie w ramach cyklu artykułów prezentowane są zarówno podejście wzorcowe [1], [2], [3], [5], [6]. [7], [8], [10], [13], [19], [20], [22] jak i bezwzorcowe [4], [9], [11], [12], [14], [15], [16], [17], [18], [21].

Zwykle w ramach wzorcowego podejścia wielomodelowego, zanim zaprezentowane zostaną metody łączenia modeli, najpierw prezentowane są najbardziej popularne klasyfikatory bazowe (modele), takie jak drzewa decyzyjne, sztuczne sieci neuronowe czy metoda *k*-najbliższych sąsiadów (por. np. Gatnar 2008; Kuncheva 2014; Zhi-Hua 2012; Rokach 2010).

Do najprostszych modeli, które z powodzeniem znajdują zastosowanie w podejściu wielomodelowym, zalicza się z pewnością metoda *k*-najbliższych sąsiadów (zob. np. Gatnar 2008). Metoda ta pozwala przydzielać obserwacje ze zbioru uczącego na podstawie K obserwacji, które leżą najbliżej obserwacji klasyfikowanej. Ostatecznie obiekt przydziela się do tej klasy, do której należy najwięcej spośród K sąsiadów obiektu klasyfikowanego. Adaptację tej metody na potrzeby danych symbolicznych zaprezentowali Malerba i in. (2001). (zob. także moja praca [3]).

Innymi prostymi modelami, które można z powodzeniem zastosować w podejściu wielomodelowym, są modele regresyjne. W przypadku danych symbolicznych mamy tu możliwość zastosowania zarówno regresji liniowej (zob. np. Diday i Noirhomme-Fraiture 2008) (praca [8] prezentuje zastosowanie regresji liniowej w podejściu wielomodelowym), jak i logitowej (de Souza i in. 2011). W pracy [21] zaprezentowałem porównanie skuteczności regresji logitowej z innymi metodami.

Jedną z ważniejszych nieparametrycznych metod, którą wykorzystuje się do budowy modeli regresyjnych, są drzewa klasyfikacyjne oraz regresyjne. Metoda ta pod nazwą rekurencyjnego podziału stosowali już Morgan i Sonquist (1963). Do celów regresyjnych oraz dyskryminacyjnych metodę tą zaproponowali w swojej pracy Breiman i in. (1984) (cyt. za

Gatnar 2008). W polskiej literaturze przedmiotu wyczerpującą publikacją w zakresie drzew regresyjnych i klasyfikacyjnych dla danych klasycznych jest praca Gatnara (2001). Drzewa decyzyjne sprawdzają się ona zarówno w przypadku danych klasycznych (por. np. Gatnar 2008) jak i symbolicznych co prezentuje na gruncie danych o znanej strukturze klas moja współautorska praca [1] oraz na gruncie danych rzeczywistych prace [19], [20], [22].

Innym z modeli, który często stosowany jest w podejściu wielomodelowym są sztuczne sieci neuronowe. Koncepcja ich budowy sięga lat czterdziestych dwudziestego wieku. Pierwszy model tego typu zaprezentowali McCulloch i Pitts (1943), natomiast mechanizm uczenia i zapamiętywania informacji zaprezentował Hebb (1949). W przypadku danych klasycznych mamy do wyboru wiele różnych rozwiązań w zakresie budowy sieci neuronowych. Natomiast w przypadku danych symbolicznych opracowano jedynie adaptację perceptronu wielowarstwowego. Jest to jednocześnie jedna najprostsza z sieci, którą dla danych klasycznych zaproponował Rosenblatt (1958). Ideę oraz zastosowanie tego modelu dla danych symbolicznych prezentują moje prace [2] i [22].

Natomiast, co do metod klasyfikacji danych symbolicznych, to w podejściu wielomodelowym zastosowanie znajdują zarówno adaptacje klasycznych metod, takie jak np. DBSCAN, *k*-medoidów, metody hierarchiczne, ale również metody klasyfikacji pojęciowej [5, 12, 13] czy nowe rozwiązania w zakresie analizy skupień, takie jak klasyfikacja spektralna (*spectral clustering*) (por. Von Luxburg 2007) oraz bazująca na przesunięciu okna średniej (*mean-shift clustering)* (zob. Derpanis 2005) – zastosowanie tych metod na gruncie podejścia wielomodelowego prezentują moje prace [14, 15].

Wśród najpopularniejszych metod budowy modeli zagregowanych w zagadnieniach dyskryminacyjnych oraz regresyjnych wyróżnia się metodę *bagging* oraz *boosting*. *Bagging*, którą zaproponował Breiman (1996), jest jedną z najbardziej znanych metod budowy modeli zagregowanych. Metoda ta polega na zbudowaniu *M* modeli bazowych na podstawie zbioru danych, z którego obiekty są losowane ze zwracaniem. Adaptację tego podejścia na potrzeby danych symbolicznych prezentują moje praca [6]. Metoda *bagging* znajduje także zastosowanie w budowie modeli zagregowanych na potrzeby klasyfikacji danych klasycznych (Leisch 1999) oraz symbolicznych co ukazują praca [13, 14, 15]. Pewne propozycje, co do stosowania tych metod zawiera także monografia Dudka (2013).

Inną popularną metodą budowy modeli zagregowanych jest *boosting*, którą zaproponowali Freund i Schapire (1995). Jej idea polega na poprawie dokładności predykcji modelu zagregowanego dzięki podwójnemu systemowi wag. Zastosowanie tego rozwiązania na potrzeby danych symbolicznych prezentuje moja praca [7].

W przypadku metod klasyfikacji danych symbolicznych do tworzenia i łączenia modeli bazowych, oprócz adaptacji metody *bagging*, można zastosować także macierz współwystąpień [Fred i Jain 2005] czy metodę zaproponowaną przez Hornika (2005).

4.3. Zakres badań

Zakres moich zainteresowań naukowych obejmuje następujące, powiązane ze sobą, obszary badawcze:

- klasyfikatory bazowe, czyli metody, które mogą podlegać łączeniu w podejściu wielomodelowym danych symbolicznych,
- metody budowy modeli zagregowanych na potrzeby dyskryminacji i regresji oraz w analizie skupień,
- ocena prezentowanych rozwiązań na podstawie zbiorów danych o znanej strukturze klas oraz rzeczywistych problemów ekonomicznych.

Proponowany tytuł osiągniecia naukowego, na który składa się prezentowany cykl artykułów wskazuje, że w znacznej mierze powinien on dotyczyć trzeciego z wymienionych obszarów badawczych. Niemniej jednak z literatury dotyczącej podejścia wielomodelowego oraz prowadzonych przeze mnie badań, wyłania się konieczność prezentowania także podstaw teoretycznych związanych zarówno z samymi klasyfikatorami bazowymi, czyli metodami, które będą podlegać łączeniu oraz metod budowy i łączenia metod na potrzeby dyskryminacji, regresji oraz analizy skupień. Stąd w prezentowanym cyklu znalazły się artykuły dotyczące modeli bazowych (klasyfikatorów bazowych) – prace [1, 2, 3, 4, 5] oraz metod budowy modeli zagregowanych – prace [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 22].

Niemniej jednak, oprócz prezentowania samej części tylko teoretycznej na gruncie analizy danych symbolicznych oraz podejścia wielomodelowego dla tego typu danych, prace realizujące dwa pierwsze cele zawierają ocenę proponowanych rozwiązań na przykładzie zbiorów danych o znanej strukturze klas i często także rzeczywistych zbiorów danych.

W przypadku prac o charakterze stricte aplikacyjnym warto wskazać na prace dotyczące obszaru Unii Europejskiej i obszaru OECD, gdzie badałem zarówno rozwój gospodarczy w ramach obszaru OECD (praca [18]), jak i innowacyjność krajów Unii (praca [17]) czy poziom wolności gospodarczej w krajach Unii (praca [21]).

Innym istotnym zagadnieniem aplikacyjnym jest z pewnością zagadnienie oceny zdolności kredytowej osób fizycznych (praca [20]) oraz ryzyko odejścia klienta (praca [19]) czy ocena pozycji produktów na rynku (praca [10]).

W celu uszczegółowienia zakresu prowadzonych przeze mnie badań związanych z przedstawionym cyklem publikacji oraz obszarami badawczymi należy wymienić:

- badania o charakterze analitycznym i symulacyjnym dotyczące klasyfikatorów bazowych na potrzeby dyskryminacji i regresji (prace [1, 2, 3, 8, 19, 20, 22]),
- badania o charakterze analitycznym i symulacyjnym dotyczące klasyfikatorów bazowych na potrzeby analizy skupień (prace [4, 5, 9, 10, 12, 15, 16, 17, 18]),
- badania o charakterze analitycznym i symulacyjnym z zakresu budowy modeli zagregowanych na potrzeby dyskryminacji i regresji (prace [6, 7, 13, 22]),
- badania o charakterze metodologicznym, analitycznym i symulacyjnym dotyczące porównywania różnych rozwiązań z zakresu podejścia wielomodelowego [1, 11, 14, 20, 22].

4.4. Teoretyczne i aplikacyjne osiągnięcia naukowe

Do osiągnięć naukowych związanych z cyklem publikacji należy zaliczyć:

A. Autorską propozycję zmodyfikowania klasyfikacji pojęciowej na potrzeby danych symbolicznych oraz zastosowanie tego podejścia w klasyfikacji wielomodelowej.

B. Inne, niż klasyfikacja pojęciowa, metody analizy danych symbolicznych, które mogą znaleźć zastosowanie w podejściu wielomodelowym.

C. Propozycję adaptacji metod *bagging* i *boosting* na potrzeby podejścia wielomodelowego danych symbolicznych oraz klasyfikację metod łączenia klasyfikacji bazowych na potrzeby analizy skupień danych symbolicznych.

D. Zastosowanie i ocena efektywności różnorodnych podejść wielomodelowych zarówno w zagadnieniach dyskryminacji, regresji, jak i klasyfikacji.

4.5. Prezentacja osiągnięć naukowych

A. Autorska propozycja zmodyfikowania klasyfikacji pojęciowej na potrzeby danych symbolicznych oraz zastosowanie tego podejścia w klasyfikacji wielomodelowej

Wśród metod klasyfikacji danych symbolicznych wyróżnia się, podobnie jak w przypadku metod klasyfikacji dla danych klasycznych, metody iteracyjno-optymalizacyjne, hierarchiczne czy wreszcie metody gęstościowe (np. DBSCAN). Niemniej jednak jako klasyfikator bazowy zastosowanie mogą znaleźć również metody klasyfikacji pojęciowej. Zgodnie z definicją

Gatnara (1998, s. 7) pojęcie jest poznawczą reprezentacją skończonej liczby wspólnych cech, które w jednakowym stopniu przysługują wszystkim reprezentantom (desygnatom) danej klasy.

Podejście bazujące na klasyfikacji pojęciowej[1] pozwala odejść od typowych metod klasyfikacji, które bazują na miarach odległości. W analizie danych symbolicznych, zanim zaproponowałem artykuł [5], do klasyfikacji pojęciowej obiektów symbolicznych opisywanych przez różne zmienne zastosować można było jedynie metodę piramid/hierarchiczną (Diday i Brito 1989).

W artykule [5] przedstawiłem ideę metody COBWEB (Fisher 1987a oraz Fisher 1987b) oraz jej modyfikacji dla danych klasycznych różnych typów, a następnie zaproponowałem na tej podstawie autorską modyfikację tej metody, która pozwala stosować ją do klasyfikacji obiektów symbolicznych opisywanych przez zmienne symboliczne różnych typów. Istotnym elementem tej modyfikacji jest całkowita użyteczność, której elementy są obliczane zależnie od typu zmiennej symbolicznej opisującej dany obiekt.

Porównywanie użyteczności metod klasyfikacji na potrzeby podejścia wielomodelowego jest niezwykle istotnym zagadnieniem, co prezentują m. in. prace: Ghaemi i in. (2009). W artykule [12] zaprezentowałem porównanie zastosowania klasyfikacji pojęciowej bazującej na metodzie hierarchicznej z dobrze znaną metodą *k*-medoidów. Do oceny wykorzystałem skorygowany indeks Randa oraz zbiory danych o znanej strukturze klas (wygenerowane w programie R) oraz dwa zbiory danych rzeczywistych. Otrzymane wyniki wskazują, że podejście wielomodelowe bazujące na klasyfikacji pojęciowej może być użytecznym narzędziem w analizę danych symbolicznych różnych typów.

Artykuł [13] stanowi swojego rodzaju połączenie pomiędzy klasyfikacją pojęciową i podejściem wielomodelowym, a dokładniej mówiąc metodą *bagging* w klasyfikacji. Wykorzystałem w nim hierarchiczną metodę klasyfikacji pojęciowej P. Brito (1989) oraz adaptację metody *bagging* zaproponowaną przez Leischa (1999). Do oceny zaproponowanego rozwiązania wykorzystałem zbiory danych o znanej strukturze klas (wygenerowane za pomocą pakietu `clusterSim` z zastosowaniem funkcji `clusterGen` (Walesiak, Dudek 2021). Dodatkowo zastosowałem zbiory danych rzeczywistych: dane dotyczące 28 modeli samochodów należących do trzech różnych segmentów (A, B, C oraz D). W przypadku sztucznych zbiorów danych otrzymane obiekty opisujące klasy pozwoliły trafnie zidentyfikować dwie (w przypadku zbioru pierwszego) oraz trzy (w przypadku zbioru drugiego) klasy. W przypadku danych rzeczywistych wyniki klasyfikacji pojęciowej

[1] Szerzej o metodach klasyfikacji pojęciowej pisze w swojej monografii Gatnar (1998).

pokrywały się z klasyfikacją pojęciową dla całego zbioru danych, a stabilność klasyfikacji, oceniona skorygowanym indeksem Randa świadczyła o relatywnie stabilnym podziale obiektów.

B. Inne, niż klasyfikacja pojęciowa, metody analizy danych symbolicznych, które mogą znaleźć zastosowanie w podejściu wielomodelowym.

W ramach podejścia wielomodelowego analizy danych symbolicznych, zarówno wzorcowego jak i bezwzorcowego, zastosowanie znaleźć mogą różnorodne metody. Podobnie, jak w przypadku danych klasycznych, ważnym zagadnieniem jest analiza i ocena efektywności poszczególnych algorytmów.

Jednym z ważniejszych algorytmów, który znajduje zastosowanie w podejściu wielomodelowym są drzewa klasyfikacyjne i regresyjne. Jedną z pierwszych propozycji sekwencyjnego podziału początkowej przestrzeni na segmenty, nazywane też regionami czy podprzestrzeniami, zastosowano w statystyce przez Morgana i Sonquista (19963). Zastosowanie tego podejścia w zagadnieniach dyskryminacyjnych i regresyjnych zawarto w pracy Berimana i in. (1984). W polskiej literaturze przedmiotu obszernie zagadnienie drzew klasyfikacyjnych i decyzyjnych dla danych klasycznych prezentuje praca Gatnara (2001).

Na potrzeby analizy danych symbolicznych opracowano jedynie adaptację drzew klasyfikacyjnych. Drzewa klasyfikacyjne dla danych symbolicznych można podzielić na drzewa klasyfikacyjne oparte na optymalnym podziale (Périnel, Lechevallier 2000, s. 245-261), warstwowe drzewa decyzyjne (*strata decision trees*) (Bravo 2000; Bravo i García-Santesmases 2000; Noirhomme i in. 2004, s. 273-283) oraz Bayesowskie drzewa decyzyjne *(Bayesian decision trees*) (Noirhomme-Fraiture i in. 2004, s. 287-294). W artykule [1] zaprezentowałem ideę drzew decyzyjnych opartych na optymalnym podziale oraz jądrowej analizy dyskryminacyjnej danych symbolicznych i następnie porównałem skuteczność tych algorytmów w identyfikacji zbiorów danych o znanej strukturze klas w sytuacji gdy zbiór danych zawiera zmienne zakłócające istniejącą strukturę klas. Do każdego zbioru danych dodawałem 2, 3, 5 oraz 10 zmiennych zakłócających. W przypadku zbiorów danych, które nie zawierały zmiennych zakłócających, jądrowa analiza dyskryminacyjna osiągała lepsze wyniki (w sensie mniejszego błędu klasyfikacji) niż drzewa decyzyjne oparte na optymalnym podziale. Wraz ze wzrostem liczby zmiennych zakłócających drzewa decyzyjne okazały się być lepszym rozwiązaniem niż jądrowa analiza dyskryminacyjna. Artykuł [20] prezentuje autorską propozycję zastosowania jednostopniowych drzew decyzyjnych (*decision stumps*) w ocenie zdolności kredytowej osób fizycznych.

Innym, istotnym z punktu widzenia podejścia wielomodelowego, algorytmem są sztuczne sieci neuronowe. Pierwszy model tego typu opracowali McCulloch i Pitts (1943), a mechanizm zapamiętywania informacji przez komórki przedstawił Hebb (1949). Najprostszą siecią neuronową, jest perceptron zaproponowany przez Rosenblatta (1958). W ramach analizy danych symbolicznych zaproponowano modyfikację perceptronu wielowarstwowego (MLP) na potrzeby obiektów opisywanych wyłącznie przez zmienne symboliczne interwałowe (Rossi i Conan-Guez 2008). Rossi i Conan-Guez (2008) zaproponowali trzy rozwiązania pozwalające na zastosowanie zmiennych symbolicznych interwałowych w ramach perceptronu wielowarstwowego: metodę ekstremów, w której zmienne symboliczne interwałowe są reprezentowane przez krańce tych zmiennych; metodę środków, w której zmienne symboliczne interwałowe są reprezentowane przez środek przedziały zmiennej; metodę próbkowania, gdzie zmienne symboliczne interwałowe zastępowane są zmienną losową o rozkładzie jednostajnym o krańcach takich, jak te w zmiennej symbolicznej. W artykule [2] wskazałem problemy, jakie mogą wynikać z zastosowania każdego z tych podejść. W części empirycznej przeanalizowałem pięć różnych zbiorów danych o znanej strukturze klas, niektóre z nich zawierały zmienne zakłócające lub/i obserwacje odstające. W każdym modeli zbiór uczący stanowiło 300 losowo dobranych obserwacji, a 100 obserwacji zbiór testowy. Do każdego z modeli zastosowałem dziewięć różnych ścieżek symulacyjnych o różnej liczbie iteracji, innej liczbie warstw ukrytych, różnej liczbie neuronów w warstwach, odrębnym współczynniku uczenia. W wyniku wstępnych analiz zdecydowałem się na zastosowanie sigmoidalnej funkcji aktywacji. Do oceny wyników osiąganych przez poszczególne podejścia zastosowałem błąd średniokwadratowy. W wyniku porównania trzech podejść wskazałem, że najlepszym rozwiązaniem jest metoda próbkowania, a najgorszym rozwiązaniem okazała się metoda środków. Generalnie perceptron wielowarstwowy dla danych symbolicznych pozawala trafnie klasyfikować zbiory danych o różnej strukturze klas (dobrze lub słabo separowalnej). Metoda ta osiąga niestety nieco gorsze wyniki, gdy w zbiorze danych mamy jednocześnie obserwacje odstające i zmienne zakłócające.

Innym, dobrze znanym klasyfikatorem bazowym jest metoda *k*-najbliższych sąsiadów. Została ona zaproponowana przez Fixa i Hodgesa (1951). Jej idea polega na klasyfikacji obserwacji do tej grupy, do której należy najwięcej spośród *K* sąsiadów (innych obserwacji) leżących najbliżej niej (zob. np. Gatnar 2008, s. 30). W artykule [3] zaprezentowałem modyfikację tej metody dla danych symbolicznych, którą zaproponowali Malerba i in. (2004). Najważniejszymi różnicami w porównaniu do wariantu dla danych klasycznych, zaliczyć

należy zastosowanie miary odległości adekwatnej dla danych symbolicznych (np. Ichino-Yaguchiego czy jednej z miar de Carvalho), a dodatkowo w procesie klasyfikacji obserwacje znajdujące się bliżej obserwacji poddawanej klasyfikacji są ważniejsze. Wynika to z wprowadzenia wag odwrotnych do odległości między obiektem klasyfikowanym, a jego sąsiadami. W części empirycznej przeanalizowałem użyteczność zaprezentowanej metody na przykładzie 3 sztucznych zbiorów danych o znanej strukturze klas, oraz dwóch zbiorach danych rzeczywistych, które przygotowali autorzy tej metody. Do każdego z modeli dodano 2, 3 lub 5 zmiennych zakłócających, a liczbę sąsiadów (10, 11 oraz 12) wybrano arbitralnie. Oceniając wyniki analiz można stwierdzić, że w przypadku, gdy mamy do czynienia ze zmiennymi zakłócającymi błąd klasyfikacji zwiększa się znacząco.

Ostatnim z klasyfikatorów bazowych, które mają duże znaczenie dla podejścia wielomodelowego są metody klasyfikacji. W ramach analizy danych symbolicznych zastosowanie znajdują zarówno metody bazujące stricte na tablicy danych symbolicznych, metody opracowane na potrzeby danych symbolicznych bazujące na macierzy odległości oraz wszystkie metody klasyczne bazujące na macierzy odległości. Do metod bazujących na tablicy danych symbolicznych zaliczają się m.in. SCLUST czy adaptacja metody *k*-średnich zaproponowana przez Verde (zob. Wilk 2010, s. 118-121 i 125). W pracach [5] zaprezentowałem autorską propozycję adaptacji metody COBWEB na potrzeby klasyfikowania obiektów symbolicznych opisywanych przez zmienne różnych typów. Natomiast prace [12 i 13] prezentują możliwość zastosowania hierarchicznej klasyfikacji pojęciowej w podejściu wielomodelowym. Wśród metod opracowanych dla danych symbolicznych, które bazują na macierzy odległości warto wskazać m. in. DCLUST, RESEARCHER, metodę deglomeracyjną Chavent (zob. Wilk 2010, s. 118-121 i 125). Oprócz tego w analizie danych symbolicznych mogą znaleźć wszystkie klasyczne metody analizy skupień, które bazują na macierzach odległości.

Praca [17] prezentuje autorską propozycję modyfikacji metody DBSCAN na potrzeby analizy danych symbolicznych. Rozwiązanie zaproponowane w tym artykule może z powodzeniem znaleźć zastosowanie w innych algorytmach klasyfikacji gęstościowej, takich jak choćby GDBSCAN (Sander i in. 1998) czy PreDeCon (Jahirabadkar i Kulkarni 2013).

W artykule [4] zaprezentowałem możliwość zastosowania klasyfikacji spektralnej na potrzeby podejścia wielomodelowego danych symbolicznych. Metoda ta nie jest nie tyle nowym algorytmem klasyfikacji danych, co nową metodą transformacji początkowej macierzy danych czy tablicy danych symbolicznych na potrzeby analizy danych. Literatura przedmiotu prezentuje wiele różnych modyfikacji klasyfikacji spektralnej m.in. w pracach Shorteed (2006)

czy Walesiaka i Dudka (2009). Istotną zaletą podejścia spektralnego jest brak założeń co do kształtu skupień, a dodatkowo radzi sobie ona znacznie lepiej z identyfikacją skupień o nietypowym kształcie. W części empirycznej artykułu porównano sześć różnych podejść w klasyfikacji danych symbolicznych z zastosowaniem podejścia spektralnego na przykładzie pięciu zbiorów danych o znanej strukturze klas. Do oceny wyników zastosowałem skorygowany indeks Randa. Najlepsze wyniki otrzymałem dla podejścia wielomodelowego, w którym podejście spektralne połączono z adaptacją metody *bagging* Leischa (1999).

C. Propozycja adaptacji metod bagging i boosting na potrzeby podejścia wielomodelowego danych symbolicznych oraz klasyfikacja metod łączenia klasyfikacji bazowych na potrzeby analizy skupień danych symbolicznych.

Metoda *bagging* zalicza się do metod, które dobierają losowo obserwacje do prób uczących. Wykorzystuje on w swojej budowie architekturę równoległą. Oznacza ona, że wykorzystywanych jest kilka zbiorów uczących, co łączy się najczęściej z losowym doborem obserwacji do prób.

Jedną z pierwszych propozycji tego typu zaproponowali Dasarathy i Sheela (1978). Zgodnie z tą propozycją każdy z modeli bazowych odpowiadał za klasyfikację obserwacji leżących w określonej części przestrzeni zmiennych. Do najbardziej znanych metod agregacji modeli bazowych jest metod bagging zaproponowana przez Breimana (1996). Polega ona na tworzeniu wielu modeli bazowych na podstawie N-elementowych prób uczących wylosowanych ze zwracaniem ze zbioru uczącego. Do łączenia wyników stosuje się zwykle metodę głosowania większościowego (dla modeli dyskryminacyjnych) lub uśredniania (dla modeli regresyjnych).

W artykule [7] zaprezentowałem możliwość adaptacji metody *bagging* na potrzeby analizowania danych symbolicznych różnego typu. Jako klasyfikator bazowy w tym przypadku posłużyła metoda *k*-najbliższych sąsiadów dla danych symbolicznych. W części empirycznej proponowane podejście zastosowano dla sztucznych zbiorów danych o znanej strukturze klas, które wygenerowano za pomocą funkcji `cluster.Gen` z pakietu `clusterSim`. W wyniku analiz okazało, się że metoda *bagging* wraz z metodą k-najbliższych sąsiadów jako klasyfikatorem bazowym pozwalają precyzyjnie identyfikować różne zbiory danych, w tym takie o słabo separowalnych klasach i wydłużonym kształcie. Niestety w przypadku zbiorów danych z obserwacjami odstającymi błąd klasyfikacji wyniósł aż 62,5%. Może to wynikać z faktu, że obserwacje odstające istotnie wpływają na pomiar odległości, a to przenosi się na prawdopodobieństwa przydzielenia obiektów do klas w metodzie k-najbliższych sąsiadów dla danych symbolicznych.

W artykule [9] przedstawiłem możliwość zastosowania metody *bagging* na potrzeby regresji liniowej danych symbolicznych. Zaprezentowałem w nim dwa podejścia (metodę środków oraz metodę środków i promieni) pozwalające wykorzystywać zmienne symboliczne interwałowe w dobrze znanej regresji liniowej, gdzie parametry są szacowane metodą najmniejszych kwadratów. Dla celów analizy utworzono od 20 do 50 modeli bazowych z zastosowaniem metody *bagging* a do łączenia wyników uśrednianie wyników. Na podstawie dwóch sztucznych i czterech rzeczywistych zbiorów danych okazało się, że metoda środków i promieni uzyskuje nieco lepsze dopasowania do danych rzeczywistych w sensie miar R^2 dla dolnych oraz górnych krańców zmiennych symbolicznych interwałowych. Niestety zarówno metoda środków jak i metoda środków i promieni, które zaproponowano w literaturze przedmiotu nie posiadają rozwiązań co do testowania założeń modelu związanych z metodą najmniejszych kwadratów.

Innymi metodami, które stosują losowy dobór obserwacji do prób uczących są m.in. *windowing* opracowany przez Quinlana (1983). Rozwiązanie to polega na tym, że próba ucząca (*window*) na podstawie której powstał model była losowym podzbiorem pierwotnego dużego zbioru danych. Innym rozwiązaniem jest *stacking* zaproponowany przez Wolperta (1992). W tym przypadku do budowy modeli bazowych sugeruje się wykorzystywanie jednoelementowego sprawdzania krzyżowego.

W ramach analizy danych symbolicznych w podejściu wielomodelowym na potrzeby analizy skupień do łączenia wyników wykorzystywać można wiele różnych podejść. Pierwszym z nich jest propozycja de Carvalho i in. (2012), gdzie podstawą do klasyfikacji są tak naprawdę różnorodne macierze odległości, które następnie są wykorzystywane do otrzymania wag obserwacji. Następnie macierze odległości oraz wektory wag są wykorzystywane do klasyfikacji obserwacji. Nie jest to więc typowe rozwiązanie podejścia wielomodelowego w klasyfikacji, gdzie zazwyczaj łączy się wyniki wielu różnych klasyfikacji bazowych. Wśród metod, które służą do łączenia wyników wielu klasyfikacji w jeden zagregowany wynik zastosowanie mogą znaleźć metody bazujące na funkcjach agregujących (*consensus functions*) albo adaptacje metody *bagging* w klasyfikacji. W ramach funkcji agregujących znajdują się takie rozwiązania jak metoda podziału hipergrafów, gdzie klasy są reprezentowane przez hiperkrawędzie (Strehl i Ghosh 2002). Podstawowym problemem jest tu znalezienie minimalnego rozcięcia. Innym rozwiązaniem jest metoda głosowania, której propozycję przedstawili m.in. Dudoit i Fridlyand (2003). Głównym celem metody głosowania jest dokonywanie permutacji etykiet klas, w taki sposób aby otrzymać jak największą zgodność pomiędzy etykietami klasy z danej metody klasyfikacji a etykietami z klasyfikacji odniesienia. W przypadku metod bazujących na teoriach informacyjnych wykorzystuje się miary informacji

wzajemnej. Innym rozwiązaniem jest rozwiązanie bazujące na metodach mieszanek, gdzie zakłada się, że wynik (etykiety klas) są modelowane jako zmienne losowe pobrane z dwóch rozkładów prawdopodobieństwa. Do łączenia wyników stosuje się tutaj metodę maksymalizacji wartości oczekiwanej (Gathemi i in 2009, s. 641).

W pracy Fredy i Jain (2005) do łączenia wyników klasyfikacji zaproponowano macierz współwystąpień (*co-clustering matrix, co-association matrix*). Macierz ta zawiera informacje ile razy obiekty *i, j* zostały zaklasyfikowane do jednej grupy w wielu klasyfikacjach bazowych. Zgodnie z ideą tego artykułu relatywnie częstsze klasyfikowanie tych obiektów do tych samych klas świadczy o tym, że powinny zostać ostatecznie zaklasyfikowane do jednej klasy. Macierz współwystąpień jest następnie stosowana jako macierz danych w innej metodzie klasyfikacji. W oryginale jest to metoda hierarchiczna, ale równie dobrze może to być inna metoda klasyfikacji.

Innymi rozwiązaniami, które mogą znaleźć rozwiązanie w przypadku metod klasyfikacji są adaptacje metody *bagging* na potrzeby analizy skupień. Propozycja Leischa (1999) polega na utworzeniu prób bootstrapowych poprzez losowanie obserwacji ze zwracaniem. Próby stanowią nowe zbiory danych, do których stosowana jest dana metoda klasyfikacji. Centra skupień otrzymanych klas stanowią zbiór danych, który poddawany jest klasyfikacji (zwykle jest to jedna z metod hierarchicznych). Obserwacje z pierwotnego zbioru danych przydzielane są do tej klasy, której zalążek znajduje się najbliżej. Modyfikacja Dudoit i Fridlyand (2003) polega na utworzeniu prób bootstrapowych, a następnie na zastosowaniu algorytmu iteracyjno-optymalizacyjnego do oryginalnego zbioru danych oraz podprób. Następnie dokonywana jest permutacja etykiet klas z prób bootstrapowych, tak aby zachodziła jak największa zgodność z etykietami z oryginalnego zbioru danych. Do otrzymania ostatecznych wyników klasyfikacji stosowane jest głosowanie majoryzacyjne. Propozycja Hornika (2005) polega na zastosowaniu klasycznego algorytmu klasyfikacyjnego do każdej z podprób i otrzymanie ostatecznego podziału poprzez wyszukanie minimum funkcji odległości.

Zastosowanie tych podejść w klasyfikacji danych symbolicznych prezentują artykuły [13], gdzie zastosowano klasyfikację pojęciową, praca [4], w którym porównano metodę Leischa, Hornika oraz Dudoit i Fridlyand na potrzeby zastosowania klasyfikacji spektralnej, artykuł [9] prezentuje różnorodne możliwe podejścia w aspekcie teoretycznym, a w części empirycznej prezentuje wyniki otrzymane dla macierzy współwystąpień. Artykuł [14] prezentuje i porównuje efektywność metody *bagging* w klasyfikacji z klasyfikacją spektralną, macierzą współwystąpień oraz klasyfikacją z przesunięciem okna średniej. W pracy [18] zastosowano metodę *bagging* na potrzeby analizy poziomu rozwoju krajów OECD.

D. Zastosowanie i ocena efektywności różnorodnych podejść wielomodelowych zarówno w zagadnieniach dyskryminacji, regresji, jak i klasyfikacji.

Praca [9] stanowi teoretyczne ujęcie podejścia wielomodelowego, gdzie prezentuję podstawowe idee i metody, które pozwalają na zastosowanie tego typu rozwiązań na potrzeby analizy danych symbolicznych, ze szczególnym uwzględnieniem zmiennych symbolicznych interwałowych. Zaprezentowałem tu podstawowe idee i algorytmy znane z podejścia wielomodelowego dla danych klasycznych. W części empirycznej dokonałem oceny przydatności macierzy współwystąpień jako narzędzia łączenia wyników wielu klasyfikacji w analizie skupień dla danych symbolicznych. W tym celu przygotowałem cztery sztuczne zbiory danych o znanej strukturze klas. Do ostatecznego wyboru liczby klas wykorzystałem tu indeksy: sylwetkowy, Bakera i Huberta, Huberta i Lewine'a i wybrałem tą liczbę klas, którą wskazywała większość z indeksów. Ponadto do oceny stabilności wyników zastosowałem skorygowany indeks Randa, a wyniki podejścia wielomodelowego porównano z pojedynczymi metodami klasyfikacji (*k*-medoidów, pojedynczego połączenia, średniej klasowej). Otrzymane wyniki wskazują, że podejście wielomodelowe uzyskuje lepsze wyniki (w sensie skorygowanego indeksu Randa) niż pojedyncze metody klasyfikacji zwłaszcza w sytuacji, gdy mamy do czynienia z nietypowymi kształtami skupień.

Kontynuacją tematyki podjętej w artykule [9] jest artykuł [11]. Zaprezentowałem w nim efektywność podejścia wielomodelowego danych symbolicznych, które bazuje na macierzy współwystąpień, w sytuacji, gdy w zbiorze danych mamy do czynienia ze zmiennymi zakłócającymi lub obserwacjami odstającymi. Dla celów badań symulacyjnych przygotowałem pięć zbiorów danych o znanej strukturze klas do których dodałem różną liczbę obserwacji odstających lub zmiennych zakłócających. Oprócz sztucznych zbiorów danych przeprowadziłem te same analizy dla rzeczywistego zbioru danych opisującego 33 modele samochodów osobowych. Dla każdego zbioru danych dokonałem klasyfikacji różnymi metodami hierarchicznymi (aglomeracyjnymi), metodą *k*-medoidów, SClust oraz deglomeracyjną metodą hierarchiczną (DIANA). W podejściu wielomodelowym zbudowałem 20 modeli, których wyniki połączyłem za pomocą macierzy współwystąpień, a ostateczną liczbę klas ustaliłem z zastosowaniem metody *k*-średnich oraz indeksu sylwetkowego do wyboru ostatecznej liczby klas. Stabilność klasyfikacji oceniłem za pomocą indeksu Randa dla modeli pojedynczych oraz uśrednionego indeksu Randa dla modeli zagregowanych. Otrzymane wyniki wskazują, że podejście wielomodelowe danych symbolicznych bazujące na macierzy współwystąpień uzyskuje lepsze wyniki (w sensie stabilności i jakości klasyfikacji) niż

pojedyncze metody klasyfikacji zarówno w przypadku sztucznych jak i rzeczywistego zbioru danych.

Dalszym rozwinięciem problematyki prezentowanej w artykułach [9 oraz 11] stanowi artykuł [14]. Głównym celem artykułu jest porównanie podejścia wielomodelowego baziującego na macierzy współwystąpień oraz adaptacji metody *bagging* Leischa podejściem spektralnym w klasyfikacji danych (*spectral clustering*), przesunięciem okna w kierunku średniej (*mean-shift clustering*). W celu porównania tych metod przygotowano pięć zbiorów danych o znanej strukturze klas do których dodawano jedną lub dwie zmienne zakłócające, a wyniki porównano z wykorzystaniem skorygowanego indeksu Randa. Dla podejścia wielomodelowego przygotowano 30 modeli, a w przypadku metody *bagging* Leischa wylosowano za każdym razem $\frac{2}{3}$ obserwacji z pierwotnego zbioru uczącego. W przypadku sztucznych zbiorów danych adaptacja metody *bagging* Leischa osiągnęła w większości przypadków osiągnęła lepsze wyniki niż pozostałe analizowane rozwiązania.

W artykule [15] zaproponowałem zastosowanie klasyfikacji spektralnej w klasyfikacji wielomodelowej danych symbolicznych. Klasyfikacja spektralna nie jest nową metodą klasyfikacji, a raczej metodą pozwalającą przekształcić początkowy zbiór danych w nowy, łatwiejszy do klasyfikacji zbiór danych (von Luxburg i in. 2015). W celu identyfikacji na jakim etapie powinna zostać zastosowana klasyfikacja spektralna – tj. czy ma być ona zastosowana dla tablicy danych symbolicznych, dla macierzy współwystąpień, czy dla zarówno dla pierwotnego zbioru danych symbolicznych jak i macierzy współwystąpień. Przeanalizowałem skuteczność klasyfikacji spektralnej dla wszystkich trzech adaptacji metody *bagging* - tj. Hornika, Leischa oraz Dudoit i Fridlyandy. W celu porównania proponowanych podejść przygotowałem pięć zbiorów danych symbolicznych o znanej strukturze klas. Dla każdego z rozwiązań procedurę klasyfikacji powtórzyłem 20 razy, a średnie wyniki indeksu Randa pozwoliły na stwierdzenie, że najlepsze wyniki osiąga metoda *bagging* Leischa, gdzie początkowa tablica danych symbolicznych została poddana klasyfikacji spektralnej. Najgorszym rozwiązaniem okazało się zastosowanie klasyfikacji spektralnej do macierzy współwystąpień.

Artykuł [22] stanowi w pewnym sensie kontynuację artykułu [15], w tym sensie, że artykuł [15] prezentował klasyfikację spektralną jako metodę przekształcenia tablicy danych symbolicznych, a artykuł [22] prezentuje zastosowanie wielowarstwowego perceptronu dla danych symbolicznych na potrzeby oceny zdolności kredytowej osób fizycznych. W części empirycznej artykułu zastosowałem zbiór danych opisujący kredytobiorców niemieckich, który

na potrzeby monografii przygotował Dudek (2013). Wyniki otrzymane dla podejścia wielomodelowego łączącego wydobywanie zmiennych z perceptronu wielowarstwowego dla danych symbolicznych z drzewem klasyfikacyjnym dla danych klasycznych porównałem z drzewem decyzyjnym dla danych symbolicznych oraz perceptronem wielowarstwowym dla danych symbolicznych. Podejście łączące wydobywanie zmiennych z wykorzystaniem perceptronu wielowarstwowego osiągnęło nieco lepsze wyniki (w sensie błędu, specyficzności, czułości i dokładności).

Do pozycjonowania produktów czy usług, czyli określania jak dany produkt czy usługa plasuje się na tle konkurentów na rynku, można zastosować wiele różnych metod analizy danych, wśród których warto wyróżnić regresję logistyczną, analizę czynnikową, analizę skupień czy wreszcie skalowanie wielowymiarowe (Walesiak 1993, s. 20-22; Zaborski 2001, s. 30)

W pracy [10] zaprezentowałem zastosowanie podejścia wielomodelowego bazującego na macierzy współwystąpień, gdzie do ostatecznej klasyfikacji zastosowano metodę hierarchizną kompletnego połączenia, oraz adaptacji metody *bagging* zaproponowanej przez Hornika (2005), w którym zastosowano 20 prób bootstrapowych losowanych ze zwracaniem i klasyfikowanych metodą *k*-medoidów, w ocenie pozycji produktów na rynku.

W części empirycznej wykorzystałem dane tyczące 28 różnych marek samochodów osobowych, o różnych parametrach użytkowych, które opisywało 10 zmiennych symbolicznych interwałowych. Marki te należały według producentów do jednego z czterech segmentów A, B, C lub D. Ostatecznie zarówno w przypadku macierzy współwystąpień i adaptacji metody *bagging* Hornika otrzymano dwie klasy obiektów podobnych – w klasie pierwszej znalazły się pojazdy z segmentów A, B oraz C, a w klasie drugiej wyłącznie pojazdy z segmentu D.

W artykule [8] zaproponowałem teoretyczne podstawy zastosowania regresji liniowej danych symbolicznych interwałowych w podejściu wielomodelowym. W celu oceny przydatności podejścia wielomodelowego w zagadnieniach regresji liniowej danych symbolicznych interwałowych przeanalizowałem dwa sztuczne zbiory danych oraz przeanalizowałem trzy rzeczywiste zbiory danych (zbiory pszenicy, dane medyczne dotyczące liczby chorób serca, oraz liczbę przestępstw na terenie USA. Pierwszy zbiór danych otrzymano agregując dane dotyczące zbiorów pszenicy w Polsce z poziomu powiatów – zmienna zależna oraz dane dotyczące nawożenia – zmienna niezależna. Zbiory danych dotyczące chorób serca oraz przestępstw pochodziły z pakietu RSDA programu R (Rodríguez 2014). Do porównania danych wykorzystano średni absolutny błąd procentowy modelu oraz przenalizowano

dopasowanie wyników do danych na podstawie współczynników R^2 (w przypadku danych symbolicznych interwałowych otrzymujemy dwa takie współczynniki, jeden dla krańców górnych, a drugi dla krańców dolnych, dla jednej zmiennej symbolicznej interwałowej). Wyniki otrzymane z zastosowaniem podejścia wielomodelowego porównano z wynikiem otrzymanym dla pojedynczego modelu regresji. Otrzymane wyniki wskazują, że podejście wielomodelowe w regresji danych symbolicznych interwałowych osiąga znacznie lepsze wyniki, w sensie miar R^2 oraz średniego absolutnego błędu procentowego. Potwierdza to użyteczność podejścia wielomodelowego w regresji danych symbolicznych interwałowych. Natomiast analiza wyników dla pojedynczych modeli pozwalają na wskazanie, że metoda środków i promieni jest nieco lepszym rozwiązaniem niż sama tylko metoda środków w przypadku zmiennych symbolicznych interwałowych.

W literaturze przedmiotu analiza i ocena innowacyjności Polski na tle innych krajów Unii Europejskiej jest poruszana przez bardzo wielu autorów. Warto tu wskazać m.in. prace Stec (2009) i Nowaka (2012), Wojtas (2013) czy praca Rynardowskiej-Kurzbauer (2015). Niemniej jednak wszystkie te prace bazują na danych klasycznych i w literaturze przedmiotu występuje tu luka w zakresie zastosowania w tym względzie danych symbolicznych. Praca [16] prezentuje zastosowanie analizy danych symbolicznych w ocenie innowacyjności 28 krajów Unii Europejskiej. Do utworzenia tablicy danych symbolicznych wykorzystano tu dane z ostatnich pięciu lat (*temporal data aggregation*) i na tej podstawie utworzono obiekty symboliczne drugiego rzędu, które opisywało 11 zmiennych symbolicznych interwałowych. Z pierwotnego zbioru zmiennych usunięto zmienne zakłócające z zastosowaniem metody HINoV (*Heuristic Identification of NOisy Variables*) w wersji zaproponowanej przez Walesiaka i Dudka (2008).

Do oceny innowacyjności wykorzystałem tu macierz współwystąpień, która zawiera informacje ile razy obiekty trafiały do tej samej klasy we wszystkich klasyfikacjach bazowych. Macierz otrzymałem na podstawie 15 różnych klasyfikacji (m.in. pam, *k*-medoidów, metody hiearchiczne). Do ostatecznego podziału obserwacji na klasy zastosowano metodę *k*-medoidów, gdzie macierz współwystąpień wykorzystano jako macierz danych. W efekcie otrzymałem cztery klasy. Podejście wielomodelowe danych symbolicznych okazało się skutecznym i efektywnym narzędziem w ocenie innowacyjności krajów UE.. W pierwszej klasie znalazło się jedenaście krajów, które mają dość wysoki, ale mocno zróżnicowany poziom zmiennych charakteryzujących innowacyjność. Poza Republiką Czeską są to kraje tzw. starej Unii. W klasie drugiej znalazło się dziesięć krajów o przeciętnym poziomie innowacyjności. Są one dość zbliżone do siebie pod względem liczby osób zatrudnionych w sektorze B+R, wielkości importu wysokich technologii, liczby przedsiębiorstw wysokich technologii oraz

zgłoszeń do Amerykańskiego Urzędu Patentowego w dziedzinie biotechnologii. W klasie trzeciej znalazły się kraje o najmniejszym poziomie innowacyjności – Portugalia, Rumunia, Słowacja, Hiszpania, Polska. Czwarta klasa to Szwecja i Wielka Brytania. Klasa ta ma najmniejszą długość przedziału dla zmiennej zatrudnionych w sektorze nowoczesnych technologii. Jednocześnie są to kraje wysoce innowacyjne.

Kontynuacją tej tematyki jest praca [17]. Zawarłem w niej propozycję zastosowania algorytmu gęstościowego DBSCAN w ocenie innowacyjności krajów Unii Europejskiej. W tym artykule do otrzymania tablicy danych symbolicznych dokonałem agregacji danych pochodzących z poziomu regionów (220 regionów) dla roku 2017 do poziomu krajów (*contemporal data aggregation*) otrzymując w ten sposób 22 obiektów symbolicznych drugiego rzędu (krajów UE) plus Norwegia, Serbia, Szwajcaria, które opisywało 18 zmiennych symbolicznych interwałowych. Ze zbioru danych usunięto niektóre kraje ze względu na braki danych, są to: Serbia, Estonia, Cypr, Łotwa, Litwa, Luksemburg, Malta i Szwajcaria.

Do budowy macierzy współwystąpień wykorzystano tu 504 różne modele otrzymane na podstawie losowego doboru zmiennych, losowego podziału zmiennych na dwie grupy, zastosowania algorytmu DBSCAN z różnymi parametrami startowymi. Do ostatecznego podziału zbioru danych wykorzystano funkcję `cluster.Sim` z pakietu `clusterSim` (Walesiak i Dudek 2021). Funkcja ta pozwala przetestować wiele różnych podziałów z zastosowaniem różnych metod klasyfikacji i indeksów jakości klasyfikacji. W tym artykule zdecydowałem się na indeks sylwetkowy. Ostateczny podział macierzy współwystąpień na cztery klasy otrzymano stosując normalizację pozycyjną, miarę odległości GDM dla danych metrycznych oraz metodę pojedynczego połączenia.

Podobne zagadnienie prezentuje także praca [18]. Pomiar, ocena oraz porównywanie rozwoju gospodarczego i społecznego krajów i regionów stanowi istotne zagadnienie w ekonomii. Do oceny rozwoju można wykorzystać wiele różnych indeksów, np. indeks lepszej jakości życia OECD, indeks rozwoju społeczno-gospodarczego krajów (HDI). Niemniej jednak indeksy te mają pewne wady i ograniczenia (por. np. Sagar, Najam 1998; McGillivray, 1991). Dlatego też istotnym zagadnieniem jest zarówno konstrukcja bardziej efektywnych indeksów oceny rozwoju krajów, jak i dokonywanie porównań pomiędzy krajami. W pracy [18] zaprezentowałem wyniki porządkowania liniowego krajów OECD ze względu na ich rozwój oraz dokonałem ich klasyfikacji z zastosowaniem podejścia wielomodelowego.

Tablica danych symbolicznych została w tym przypadku utworzona na podstawie danych pobranych z Banku Światowego o 30 krajach z grupy OCED, które opisywało 19 zmiennych symbolicznych interwałowych. W wyniku porządkowania liniowego zidentyfikowałem kraje

o najwyższym (Islandia, Norwegia i Szwecja) i najniższym (Grecja, Bułgaria, Rumunia) poziomie rozwoju. Natomiast zastosowanie skalowania dwustopniowego (dwuetapowego) skalowania wielowymiarowego, w którym wykorzystywane jest skalowanie wielowymiarowe (Walesiak 2016 oraz 2017), pozwoliło na graficzną prezentację wyników na mapie percepcyjnej i identyfikację krajów o z zbliżonym poziomie rozwoju, które osiągnęły ten poziom dzięki innym czynnikom (wartościom zmiennych). W analizie skupień zastosowano metodę *k*-medoidów (ze znormalizowaną miarą odległości Ichino-Yaguchiego) oraz podejście wielomodelowe (adaptacja metody bagging Leischa oraz macierz współwystąpień). Zarówno podejście wielomodelowe, jak i metoda k-medoidów pozwoliły na odkrycie dwóch klas, w których znalazły się te same obiekty. Przewagę podejścia wielomodelowego nad pojedynczą metodą klasyfikacji wskazuje tu nieco wyższy poziom skorygowanego indeksu Randa.

W pracy [19] zaproponowałem podejście hybrydowe, gdzie połączyłem wspólnie z dr Anetą Rybicką *conjoint analysis* z drzewem klasyfikacyjnym dla danych symbolicznych. Głównym celem artykułu była identyfikacja czynników decydujących o odejściu klienta na przykładzie Polskiego rynku telefonów komórkowych. Zwykle w badaniach dotyczących odejścia klienta analizowane są głównie czynniki związane z lojalnością klienta (np. Śmiatacz 2012; Burez i Van den Poel 2009), natomiast pomijane jest znaczenie preferencji klienta. Artykuł ten stanowi współautorską propozycję uzupełnienia tej luki.

Tablicę danych symbolicznych tworzą w tym przypadku wyniki badania ankietowego obrazujące skłonność 109 respondentów do odejścia z sieci telefonii komórkowej (zmienne symboliczne interwałowe). Ci sami respondenci dokonali oceny 17 profilów prezentujących oferty wiodących operatorów na Polskim rynku. W wyniku połączenia wyników otrzymanych z obydwu podejść – symbolicznego oraz *conjoint analysis* – otrzymaliśmy drzewo decyzyjne, które jednocześnie obejmowało zagadnienia związane z ryzykiem odejścia klienta, jak i jego preferencjami wobec ofert. Najważniejszym czynnikiem różnicującym respondentów okazało się stwierdzenie „Korzystam z usług obecnego operatora ponieważ są dla mnie najlepszym wyborem", natomiast użyteczności cząstkowe, otrzymane z *conjoint analysis* pozwoliły zidentyfikować czynniki ryzyka odejścia klienta – są to miesięczne opłaty oraz marka operatora komórkowego. Połączenie obydwu metod pozwoliło na pogłębioną analizę i ocenę czynników, które mogą decydować o odejściu klienta, a tym samym pozwalają operatorom przeciwdziałać temu zjawisku.

Artykuł [20] prezentuje natomiast autorską propozycję zastosowania jednostopniowych drzew decyzyjnych (*decision stumps*) w podejściu wielomodelowym danych symbolicznych na potrzeby oceny zdolności kredytowej osób fizycznych. W artykule oprócz podejścia

wielomodelowego, w którym zastosowano jednostopniowe drzewa decyzyjne, do oceny zdolności kredytowej wykorzystano także drzewa decyzyjne oparte na optymalnym podziale, jądrową analizę dyskryminacyjną oraz perceptron wielowarstwowy. Modele te porównano zarówno w przypadku pojedynczego modelu, jak i podejścia wielomodelowego (50 modeli dla drzew decyzyjnych oraz jądrowej analizy dyskryminacyjnej, a 20 modeli dla perceptronu wielowarstwowego). Podejście wielomodelowe bazujące na jednostopniowych drzewach decyzyjnych jest dobrym narzędziem do oceny zdolności kredytowej osób fizycznych, podstawowe parametry służące do oceny modeli, takie jak błąd modelu, wskazują że podejście to daje zbliżone rezultaty do podejścia wielomodelowego w którym zastosowano drzewa decyzyjne oparte optymalnym podziale. Istotnym problemem w przypadku jednostopniowych drzew decyzyjnych jest ograniczona możliwość interpretacji wyników pojedynczego modelu.

Artykuł [21] prezentuje klasyfikację wielomodelową oraz porządkowania liniowe 27 krajów Europy pod względem ich wolności gospodarczej. Wolność gospodarcza w ogólnym ujęciu oznacza zdolność społeczeństwa danego kraju do podejmowania działalności gospodarczej. W literaturze przedmiotu zaproponowano wiele różnych indeksów (miar agregatowych) pozwalających ocenić wolność gospodarczą kraju, m.in. indeks wolności gospodarczej Heritage Foundation czy miernik wolności gospodarczej Instytutu Frasera.

W artykule zbiór danych stanowiło 27 wybranych krajów Europy (plus wzorzec oraz antywzorzec w porządkowaniu liniowym), które opisywało 12 zmiennych symbolicznych interwałowych otrzymanych dzięki połączeniu informacji z lat 2016-2019 (*temporal data aggregation*). W przypadku klasyfikacji wielomodelowej zastosowano tu adaptację metody *bagging* zaproponowaną przez Leischa (1999). Na potrzeby porządkowania liniowego obiektów symbolicznych obliczono sześć różnych miar odległości, które następnie połączono w jedną (zagregowaną) macierz odległości korzystając z propozycji Melssena i in. (2006). Na podstawie otrzymanej macierzy odległości przeprowadzono porządkowanie liniowe z wizualizacją wyników. Krajami o największym poziomie wolności gospodarczej są Wielka Brytania, Dania i Szwecja. W wyniku klasyfikacji wielomodelowej z zastosowaniem modyfikacji algorytmu *bagging* Leischa (25 podzbiorów danych, z których każdy zawierał 10 obiektów) otrzymano trzy klasy, dla których indeks sylwetkowy wyniósł 0,6287997. Dodatkowo dokonałem identyfikacji, które zmienne decydują o przynależności do klas z zastosowaniem drzew klasyfikacyjnych dla danych symbolicznych opartych na optymalnym podziale. Najistotniejszymi zmiennymi okazały się związana z kapitałem oraz swoboda pracy.

6. Literatura

Asmis E. (1984), *Epicurus' scientific method*. Cornell University Press, Ithaca.

Burez J., Van den Poel D. (2009), *Handling class imbalance in customer churn prediction*. Expert Systems with Applications, 36(3), s. 4626-4636.

De Carvalho F., Lechevallier Y., de Melo F., (2012), *Partitioning hard clustering algorithms based on multiple dissimilarity matrices*. Pattern Recognition, 45(1), s. 447-464.

Bravo C. (2000), *Strata decision trees*. [W:] H.-H. Bock, E. Diday (red.), *Analysis of Symbolic Data. Exploratory methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Belin-Heidelberg.

Bravo C., García-Santesmases J. (2000), *Symbolic object description of strata by decision trees*. Computational Statistics. Physica, 15(1), s. 13-24.

Diday E., Brito P. (1989), *Symbolic cluster analysis*. [W:] O. Opitz (red.), *Conceptual and Numerical Analysis of Data,* Springer-Verlag, Berlin-Heidelberg, s. 45-84.

Diday E., Noirhomme-Fraiture M. (2008), *Symbolic data analysis and SODAS software*. Wiley, New York.

Derpanis K. (2005), *Mean shift clustering*. Lecture Notes, 32, s. 1-4.

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

Efron B. (1979*), Bootstrap methods: another look at the jackknife*. The Annals of Statistics, 7(1), s. 1-26.

Fisher D. (1987a), *Knowledge acquisition via incremental conceptual clustering*. Machine Learning, vol. 2, s. 139–172.

Fisher D. (1987b), *Knowledge acquisition via incremental conceptual clustering*. Technical Report no. 87-22, University of California, Irvine.

Fix E., Hodges J. (1951), *Discriminatory analysis, nonparametric discrimination: Consistency properties*. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas.

Fred A., Jain A.(2005), *Combining multiple clustering using evidence accumulation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, s. 835-850.

Freund Y., Schapire R. (1995), *A decision-theoretic generalization of on-line learning and an application to boosting*. [W:] Proceedings of the Second European Conference on Computational Learning Theory, Springer-Verlag, s. 23-27.

Gathemi R., Sulaiman N., Ibrahim H., Mustapha N. (2009), *A survey: Clustering ensemble techniques*. Proceedings of World Academy of Science, Engineering and Technology, vol. 38, s. 636-645.

Gatnar E. (1993), *Modelowanie jakościowe zjawisk ekonomicznych*. Akademia Ekonomiczna im. Karola Adamieckiego w Katowicach, rozprawa doktorska (maszynopis powielony).

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

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

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

Hebb D. (1949), *The organization of behavior*. Wiley, New York.

Hornik K. (2005), *A CLUE for CLUster ensembles*. Journal of Statistical Software, 14(1), 1-25.

Jahirabadkar S., Kulkarni P. (2013), *Clustering for high dimensional data: density based subspace clustering algorithms*. International Journal of Computer Applications, 63(20).

Kuncheva L. (2014), *Combining pattern classifiers: Methods and Algorithms*. Wiley, Hoboken.

Leisch F. (1999), *Bagged clustering*. Adaptive Information Systems and Modeling in Economics and Management Science, Working Papers, SFB, 51.

Malerba D., Esposito F., D'Amato C., Appice A. (2004), *K-nearest neighbor classification for symbolic objects.* [W:] P. Brito, M. Noirhomme-Fraiture (red.), *Symbolic and spatial data analysis: mining complex data structures*. University of Pisa, Pisa, s. 19-30.

McCulloch W., Pitts W. (1943), *A logical calculus of ideas imminent in nervous activity*. Bulletin of the Mathematical Biophysics, 5, s. 115-133.

McGillivray, M. (1991). *The human development index: yet another redundant composite development indicator?* World Development, 19 (10), s. 1461‑1468.

Morgan J., Sonquist J. (1963), *Problems in analysis of survey data: a proposal*. Journal of the American Statistical Association, 58, s. 417-434.

Nowak P. (2012), *Poziom innowacyjności polskiej gospodarki na tle krajów UE*, Prace Komisji Geografii Przemysłu, nr 19, s. 153-168.

Noirhomme-Fraiture M. (red.) (2004), *User manual for SODAS 2 software. Software Report*. Analysis System of Symbolic Official Data, Project no. IST-2000-25161.

Polikar R. (2006), *Ensemble based systems in decision making*. IEEE Circuits and systems magazine, 6(3), s. 21-45.

Périnel E., Lechevallier Y. (2000), *Symbolic discrimination rules*. [W:] H.-H. Bock, E. Diday (red.), *Analysis of Symbolic Data. Exploratory methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Belin-Heidelberg.

Rokach L. (2010), *Pattern classification using ensemble methods*. World Scientific, Singapore.

Rosenblatt M. (1958), *The perceptron: A probabilistic model for information storage and organization in the brain*. Psychological Review, 65(6), s. 386-408.

Rodríguez O. (2014), *The `RSDA` package for R software*. www.r-project.org.

Rossi F., Conan-Guez B. (2008). *Multi-layer perceptrons and symbolic data*. [W:] E. Diday, M. Noirhomme-Fraiture, *Symbolic Data Analysis and the SODAS Software*, Wiley, New York, s. 373-391.

Rynardowska-Kurzbauer J. (2015), *Innowacyjność wybranych krajów Europy Środkowo-Wschodniej*, Zeszyty Naukowe Politechniki Śląskiej, seria „Organizacja i Zarządzanie", nr 86, s. 93-101.

Sagar A., Najam A. (1998), *The human development index: a critical review*. Ecological economics, 25 (3), s. 249‑264.

Sander J., Ester M., Kriegel P., Xu X. (1998), *Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications*. Data mining and knowledge discovery, 2(2), s. 169-194.

Shorteed S., (2006), *Learning in spectral clustering*. Univeristy of Washington, rozprawa doktorska (maszynopis powielony).

de Souza R., Queiroz D., Cysneiros F. (2011), *Logistic regression-based pattern classifiers for symbolic interval data*. *Pattern Analysis and Applications, 14*(3), s. 273-282.

Stec M. (2009), *Innowacyjność krajów Unii Europejskiej*, Gospodarka Narodowa, 11-12, s. 45-65.

Strehl A., Ghosh H. (2002),. *Cluster ensembles – A knowledge reuse framework for combining multiple partitions*. Journal of Machine Learning Research, 3, p. 583-618.

Śmiatacz, K. (2012), *Badanie satysfakcji klientów na przykładzie rynku usług telefonii komórkowej w Polsce*. Wydawnictwo Uczelniane Uniwersytetu Technologiczno-Przyrodniczego w Bydgoszczy.

Vermaas P. (2014), *Design theories, models and their testing: on the scientific status of design research*. [W:] *An anthology of theories and models of design*. Springer, London, s. 47-66.

Vega-Pons, S. Ruiz-Shulcloper J. (2011). *A survey of clustering ensemble algorithms*. International Journal of Pattern Recognition and Artificial Intelligence, 25(03), s. 337-372.

Von Luxburg U. (2007), *A tutorial on spectral clustering*. Statistics and computing, 17(4), s. 395-416.

Walesiak M. (1993), *Statystyczna analiza wielowymiarowa w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej im. O. Langego we Wrocławiu, Wrocław.

Walesiak M., Dudek A. (2008), *Identification of Noisy Variables for Nonmetric and Symbolic Data in Cluster Analysis*, [W:] C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (red.), *Data Analysis, Machine Learning and Applications*, Springer-Verlag, Berlin-Heidelberg, s. 85-92.

Walesiak M., Dudek A. (2009), *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 84, s. 9-19.

Walesiak M. (2016), *Visualization of linear ordering results for metric data with the application of multidimensional scaling*. Ekonometria, (52), s. 9-21.

Walesiak M. (2017),*Wizualizacja wyników porządkowania liniowego dla danych porządkowych z wykorzystaniem skalowania wielowymiarowego*. Przegląd Statystyczny, 64(1), s. 5-20.

Walesiak M., Dudek A. (2021), *The* `clusterSim` *package for R software*. www.r-project.org.

Wojtas M. (2013), *Innowacyjność polskiej gospodarki na tle krajów Unii Europejskiej*, Zeszyty Naukowe Uniwersytetu Szczecińskiego, nr 756, seria „Finanse, Rynki Finansowe, Ubezpieczenia", nr 57, s. 605-617.

Wilk J. (2010), *Problemy segmentacji rynku z wykorzystaniem metod klasycznych i symbolicznych*. Uniwersytet Ekonomiczny we Wrocławiu, rozprawa doktorska (maszynopis powielony).

Zaborski A. (2001), *Skalowanie wielowymiarowe w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej im. O. Langego we Wrocławiu, Wrocław.

Zhi-Hua Z. (2012), *Ensemble methods. Foundations and algorithms*. CRC Press, Boca Raton.

Marcin Pełka

Jelenia Góra, 5 stycznia 2022 Marcin Pełka