INFORMATYKA EKONOMICZNA BUSINESS INFORMATICS 4(46) • 2017
ISSN 1507-3858
e-ISSN 2450-0003

Iwona Chomiak-Orsa, Filip Wójcik
Uniwersytet Ekonomiczny we Wrocławiu
e-mails: iwona.chomiak@ue.wroc.pl; filip.wojcik@ue.wroc.pl

WSPOMAGANIE PROCESÓW DECYZYJNYCH PRZEZ WYKORZYSTANIE ALGORYTMÓW REGUŁOWYCH WYZNACZAJĄCYCH ISTOTNOŚĆ ATRYBUTÓW

DETECTING BUSINESS-RELEVANT ATTRIBUTES IN RULE-BASED CLASSIFICATION

DOI: 10.15611/ie.2017.4.05
JEL Classification: C44, C88, D81

Streszczenie: Nowoczesne systemy wspomagania decyzji biznesowych korzystają niejednokrotnie z uczenia maszynowego i sztucznej inteligencji do rozwiązywania skomplikowanych problemów. Jednym z nich jest klasyfikacja postrzegana w tym kontekście jako przyporządkowywanie obserwacji (obiektów) określonym kategoriom. Wśród wielu metod umożliwiających osiągnięcie tego celu znajdują się algorytmy regułowe, które poza wspomaganiem decyzji pozwalają zaobserwować korelacje wewnątrz wolumenów danych. Ma to szczególne znaczenie w przypadku decyzji uwzględniających duże wolumeny danych. Procedury te napotykają jednak problemy w przypadku silnego zaburzenia proporcji kategorii lub poszczególnych atrybutów. Odpowiedzią na to wyzwanie może być skuteczna metoda wyboru cech istotnych. W artykule wykorzystano jedną z odmian testu permutacyjnego. Jako przykład zastosowania biznesowego omówione zostało wykorzystanie algorytmu RIPPER użytego do analizy wiarygodności kredytowej klientów instytucji finansowej.

Słowa kluczowe: uczenie maszynowe, sztuczna inteligencja, wspomaganie decyzji, klasyfikacja, systemy regułowe, analiza danych.

Summary: Modern decision support systems make use of machine learning and artificial intelligence to solve complicated problems. One of them is classification, understood in this context as assigning objects to categories. Amongst many methods to achieve this goal, rule-based systems pay special attention, because they provide an end-user not only with direct answers to a given problem, but also produce useful insights into correlations present in a dataset. In this article new method has been proposed − application and modification of Leo Breiman's original *Random Forest* solution combined with backwards elimination (known from classic regression) − and tested on real credit decisions dataset. Differences in classification metrics between base and augmented classifier were checked using cross-validation testing, and statistical significance. The article concludes with further research suggestions.

Keywords: machine learning, artificial intelligence, decision support, classification, rule based systems, data science.

1. Wstęp

Kluczowymi elementami zarządzania każdym przedsiębiorstwem jest nie tylko nadzór nad realizacją procesów biznesowych, ale przede wszystkim rozwiązywanie problemów dobrze, jak również słabo ustrukturalizowanych [Surma 2009, s. 14]. Bez względu na ich charakterystykę każda decyzja zarządcza musi być poparta odpowiednią informacją pozwalającą na wybór optymalnego (w danych warunkach) rozwiązania [Kisielecki 2008]. W dobie nadmiaru informacji i ich redundancji kluczowe staje się identyfikowanie najistotniejszych atrybutów, mających wpływ na podejmowane działania. Przy wzrastających ciągle wolumenach danych wyznaczanie istotności atrybutów nabiera szczególnego znaczenia, ponieważ błąd decyzji może być obarczony znacznymi kosztami zarówno dla organizacji, jak i jej klienta. Klasycznym przykładem mogą być decyzje dotyczące oceny zdolności kredytowej klientów banku. Wydanie decyzji pozytywnej wobec osoby niemającej wystarczających środków naraża wszystkie strony stosunku prawnego na potencjalne konsekwencje – prawne lub finansowe. W odwrotnej sytuacji odmowa udzielenia kredytu osobie spełniającej wymagania może (w najgorszym przypadku) doprowadzić do utraty klienta lub przedłużyć proces.

Inteligentne systemy wspomagania decyzji mogą ułatwić to zadanie przez wskazywanie korelacji i asocjacji łączących ze sobą poszczególne składniki analizowanego wolumenu. Niemniej jednak zgodnie z założeniem *klątwy wymiarowości* [Flach 2012, s. 243] im więcej dostępnych do wykorzystania cech/atrybutów, tym mniej wiarygodne stają się prognozy. Niezwykłej wagi nabiera zatem zapewnienie odpowiedniej jakości informacji zarządczych, wspomagających procesy podejmowania decyzji na wszystkich szczeblach odpowiedzialności – od taktycznego począwszy, na strategicznym skończywszy [Surma 2009, s. 15].

W niniejszym artykule zaproponowano algorytm selekcji istotnych atrybutów predykcyjnych, posługując się studium przypadku bankowych decyzji kredytowych. Zaprezentowany proces decyzyjny stanowi klasyczny przykład zastosowania sztucznej inteligencji do wspomagania procesów decyzyjnych, obarczonych ryzykiem [Surma 2009, s. 50].Wykorzystana metoda klasyfikacji regułowej, z jednej strony, zapewnia biznesowym odbiorcom końcowym pełen wgląd w wyniki i strukturę modelu, a z drugiej, jest podatna na odchylenia w rozkładzie wartości, co może prowadzić do fałszywych wniosków [Aggarwal 2015, s. 136]. Proponowany algorytm wyznaczania tzw. permutacyjnej istotności stanowi próbę zaadresowania tego problemu.

2. Charakterystyka systemów wspomagania decyzji biznesowych

Komputerowe wspomaganie decyzji biznesowych stanowi jedno z najważniejszych zastosowań systemów typu *business intelligence*. Termin ten, pochodzący z lat 80., dotyczy wielu działań mających na celu usprawnienie procesów wewnętrznych organizacji i ułatwienie decydentom działań operacyjnych [Surma 2009]. Tradycyjnie

zalicza się do tej kategorii systemy ekspertowe, opierające się na skodyfikowanej w postaci reguł wiedzy specjalistów dziedzinowych [Olszak, Mach-Król 2016]. Współczesne możliwości w zakresie indukcyjnego gromadzenia doświadczenia, automatycznego wyciągania wniosków i aplikowania posiadanej wiedzy do rozwiązywania nowych problemów sprawiają, iż także uczenie maszynowe oraz rozpoznawanie wzorców stają się nieodłącznymi elementami BI [Chomiak-Orsa, Mrozek 2017]. W połączeniu z odpowiednią architekturą hurtowni danych tworzą narzędzia odkrywania nowej wiedzy, a także wsparcie dla wszystkich szczebli zarządzania organizacji [Olszak 2012]. Do najistotniejszych kategorii problemów rozwiązywanych przez sztuczne inteligencje należą:

Klasyfikacja – wieloklasowa lub binarna, będąca przedmiotem niniejszego artykułu.

Grupowanie – polegające na łączeniu elementów zbliżonych pod względem badanych cech.

Regresja – przewidywanie wartości numerycznych.

Z biznesowego punktu widzenia klasyfikacja odgrywa bardzo ważną rolę – buduje uproszczone modele opisujące określone zjawiska i stanowiące uogólnienie wiedzy, opartej na napotkanych przypadkach, zgromadzonej przez system. W zależności od wykorzystanej techniki lub algorytmu modele klasyfikacyjne mogą poddawać się inspekcji i analizie, dostarczając dodatkowych informacji o przesłankach stojących za podejmowanymi decyzjami [Morzy 2013].

Klasyfikacja, obok grupowania, regresji i rozpoznawania wzorców, należy do najpowszechniejszych zadań wykonywanych przez systemy uczenia maszynowego [Morzy 2013]. Stanowi także podkategorię szerszej metodologii, zwanej uczeniem z nauczycielem (lub uczeniem nadzorowanym), w którym obecne są dwa zbiory danych – treningowy, służący do nauczania, oraz testowy, używany w celu weryfikacji rezultatów. Specjalna funkcja, określana jako wyrocznia lub nauczyciel, przyznaje oceny będące odzwierciedleniem jakości postępów dokonywanych przez algorytm.

Mając daną tzw. przestrzeń przykładów (zwaną przez niektórych autorów dziedziną, opisującą obiekty/rekordy i oznaczaną jako X, oraz zestaw etykiet docelowych (kategorii), do których one należą – oznaczaną jako $\mathcal{L}$, system klasyfikacyjny stara się odnaleźć funkcję mapującą obiekty na kategorie. Nosi ona nazwę pojęcia docelowego [Flach 2012, s. 50-51].

3. Charakterystyka i zastosowania systemów regułowych

Systemy regułowe należą do jednych z najwcześniejszych algorytmów uczenia maszynowego, najbliżej spokrewnionych z klasycznymi systemami eksperckimi [Mitchell 1997, s. 20-21]. Ich głównym zadaniem jest wyszukiwanie reguł, zdolnych objaśniać korelacje i zależności zachodzące w zbiorach danych – mapując właściwości przykładów na etykiety kategorii. Algorytmy te tworzą bardzo rozległą rodzi-

nę, o wewnętrznym zróżnicowaniu, zależnym od konkretnej implementacji. Łączą je jednak pewne wspólne cechy charakterystyczne:

- **Czytelność struktury** – każdy model składa się z reguł mających postać warunkową: *jeżeli* [**poprzednik]**, *to* [**następnik]**.

Z punktu widzenia odbiorcy biznesowego czytelność struktury modelu jest jedną z jego najistotniejszych właściwości. W przeciwieństwie do tak zwanych modeli czarnoskrzynkowych [Hastie, Tibshirani, Friedman 2009, s. 352-53], tzn. takich, których struktura pozostaje niezrozumiała dla człowieka, zapewniają one dokładny wgląd w elementy składające się na proces decyzyjny.

- **Opisowe własności statystyczne –** ponownie, w przeciwieństwie do wzmiankowanych systemów czarnoskrzynkowych, każda z reguł opisana jest szeregiem miar statystycznych, wskazujących na stopień jej powiązania zarówno z klasą docelową, jak i dopasowaniem. W połączeniu z punktem poprzednim daje to odbiorcy biznesowemu możliwość zapoznania się z jakością każdej reguły, wpływającej na proces podejmowania decyzji przez automat, podobnie jak miało to miejsce w klasycznych systemach eksperckich.

Jednym z najbardziej znanych algorytmów regułowych jest RIPPER (*Repeated Incremental Prunning to Produce Error Reduction*), zaproponowany po raz pierwszy w 1995 r. [Cohen 1995]. Generuje on reguły decyzyjne dla każdej klasy z osobna w iteracjach. Wolumen danych jest dzielony na tzw. konstrukcyjny, pozwalający komplikować reguły w miarę nabierania doświadczenia, oraz tzw. walidacyjny, kontrolujący złożoność całości.

Procedura ta została wybrana do analizy z kilku względów. Po pierwsze, reprezentuje grupę algorytmów białoskrzynkowych, opartych na regułach – a więc czytelnych i zrozumiałych dla człowieka. Co więcej – należy ona do najtrafniejszych w swojej kategorii. Jednocześnie jest mniej skuteczna niż np. sieci neuronowe lub drzewa decyzyjne. Po drugie, RIPPER jest wrażliwy na niezrównoważone proporcje klas w zbiorach danych. Po trzecie wreszcie – jest stosunkowo wolny, gdy musi przetworzyć wiele atrybutów iteracyjnie. Taki zestaw cech pozwala zaprezentować metody optymalizacji, będące przedmiotem niniejszej publikacji.

4. Zidentyfikowany problem badawczy oraz proponowana metoda ustalania istotności atrybutów

Sam fakt istnienia skutecznego klasyfikatora nie oznacza, iż jest on w stanie uzyskać zadowalające rezultaty za każdym razem. Kluczowym elementem pozostaje, jak wspomina C. Shearer w swojej publikacji z 2000 r., stanowiącej podstawę metodologii analitycznej CRISP-DM [Shearer 2000], odpowiedni wybór i selekcja atrybutów wykorzystywanych przez klasyfikator. Z punktu widzenia wspierania zarządczych decyzji biznesowych obarczonych ryzykiem nabiera to szczególnego znaczenia. W środowisku redundancji informacyjnej menedżerowie bardzo często nie są świadomi istnienia pewnych istotnych implikacji, a nadmiar danych nie po-

zwala zidentyfikować korelacji iluzorycznych lub wręcz błędnych [Surma 2009, s. 14-15]. Jednym z głównych zadań systemów doradczych oraz EIS (*Executive Information System*) jest ułatwianie kierownictwu podejmowania decyzji oraz dostarczanie wartościowych informacji [Kisielecki 2008, s. 293-294]. Nie jest to możliwe, gdy system nie potrafi sobie sam poradzić z filtracją atrybutów nieistotnych.

W odpowiedzi na problemy wzmiankowane w punkcie poprzednim proponowana jest metoda wyznaczania atrybutów istotnych dla klasyfikacji algorytmem RIPPER, zwana dalej metodą permutacyjnej istotności dla trafności zrównoważonej (BART – *Balanced Accuracy Permutaiton Test*). Stanowi ona modyfikację oryginalnie zaproponowanego w 1994 r. testu permutacyjnego do sprawdzania pojedynczej zmiennej losowej pod kątem zgodności z zadanym rozkładem [Good 1994]. Wzmiankowana metoda była następnie rozwijana, głównie pod kątem optymalizacji lasów losowych (*random forest*) [Breiman 2001] oraz pojedynczych drzew decyzyjnych [Eibe, Witten 1998]. Inne prace [Ojala, Garriga 2010] traktują testy permutacyjne jako stabilniejszą, w porównaniu np. z walidacją krzyżową, metodę wyznaczania rzeczywistej trafności ogólnej klasyfikatora. W niniejszym artykule proponowane jest jego wykorzystanie w charakterze sposobu odróżnienia cech istotnych w warunkach silnego zaburzenia proporcji klas. Jest ono oparte na algorytmie zaproponowanym w 2004 r. [Radivojac i in. 2004], przeznaczonym dla bardziej stabilnych zbiorów danych.

Metoda testów permutacyjnych dla klasyfikatorów opiera się na założeniu, iż jedyną wiarygodną metodą zbierania informacji, czy cecha wpływa pozytywnie czy negatywnie na trafność, jest wyliczenie tzw. empirycznych *p*-wartości (*empirical p-value*), będących odpowiednikiem *p*-wartości klasycznych testów statystycznych [Ojala, Garriga 2010].

W tym kontekście empiryczna *p*-wartość jest proporcją przypadków, gdzie klasyfikator operujący na losowej permutacji cech uzyskał wynik lepszy (niższy błąd) niż dla cech oryginalnych, co stanowi anomalię. *A contrario,* atrybuty najbardziej istotne to te, dla których wyniki po permutacji pogarszały się, a zatem *p*-wartość była niska. Statystyczna istotność jest ustalana na zadanym poziome α, podobnie jak ma to miejsce przy innych testach. W pracy *Feature Selection Filters Based on the Permutation Test* [Radivojac i in. 2004] zaproponowano dokonywanie permutacji cech w obrębie każdej klasy z osobna i badanie, jak zmieni się trafność klasyfikatora, opierając się na mierze zysku informacyjnego lub statystyce chi-kwadrat. W celu uzyskania lepszej stabilności wyników procedurę przeprowadzano w ramach *k*-krotnej walidacji krzyżowej, uśredniając wyniki.

Na tej podstawie można zidentyfikować, które zmienne wpływają na jakość klasyfikacji. Atrybuty uznane za nieistotne można usunąć, zmniejszając tym samym wymiarowość danych, przyspieszając czas działania algorytmu i podnosząc trafność prognozy.

5. Metodologia oraz wyniki badań

W celu weryfikacji zaproponowanego podejścia autorów przeprowadzono badania empiryczne z wykorzystaniem rzeczywistych zbiorów danych. Zasadniczym ich celem było wykazanie przydatności wzmiankowanego podejścia w biznesowych zastosowaniach, mogących wspierać procesy podejmowania decyzji. Stwierdzenie, które atrybuty wnoszą rzeczywisty wkład w trafność analiz, ma kluczowe znaczenie, zgodnie z metodologią badań analitycznych CRISP-DM.

Jako reprezentatywny przykład wybrano zestaw „German Credit Data", dostarczony przez profesora Hansa Hoffmana (z Instytutu Statystyki i Ekonometrii Uniwersytetu w Hamburgu), dostępny w publicznym repozytorium naukowo-badawczym UCI Machine Learnig Repository[1] – platformie przeznaczonej do dzielenia się zbiorami danych interesującymi z punktu widzenia uczenia maszynowego, statystyki i szeroko rozumianej analizy danych. Opisuje on binarną ocenę zdolności klientów niemieckiej instytucji finansowej (chcącej zachować anonimowość) do otrzymania karty kredytowej[2]. Każda osoba opisana jest szeregiem atrybutów powiązanych z jej statusem materialnym, stanem posiadania, historią zobowiązań itd. Dane zostały zanonimizowane, aby uniemożliwić identyfikację podmiotów.

Testy przeprowadzono metodą *k*-krotnej walidacji krzyżowej, polegającej na *k*-krotnym, różnorodnym podziale zbioru uczącego na część trenująca i testową. Zapewnia to mniejszą wariancję wyników i lepsze przybliżenie sprawności rzeczywistej [Flach 2012, s. 349]. Zgodnie z proponowanym algorytmem metodą permutacyjnego testu istotności z empirycznymi *p*-wartościami wyselekcjonowano cechy negatywnie wpływające na trafność. Połączono je w podzbiory (po 2, po 3 itd.), sprawdzając, czy ich łączne usunięcie wpłynie pozytywnie na trafność predykcji na zbiorze testowym.

Następnie, chcąc porównać ze sobą dwa klasyfikatory, przeprowadzono test istotności statystycznej *t*-Studenta dla różnicy średniej trafności zrównoważonej na zadanym poziomie istotności $\alpha = 0{,}05$. Hipotezy testu byłyby następujące:

- **Hipoteza zerowa**: różnica w metryce trafności zrównoważonej pomiędzy klasyfikatorami wynosi zero.
- **Hipoteza alternatywna**: wartość metryki trafności zrównoważonej dla klasyfikatora zredukowanego jest większa od wartości miary dla klasyfikatora bazowego.

Podzbiór atrybutów uznawany jest za nieistotny w klasyfikacji, jeśli różnica średnich trafności zrównoważonych jest istotna statystycznie i wyższa dla klasyfikatora po jego usunięciu – oznacza to, iż atrybuty poddane przypadkowej permutacji dają lepsze wyniki niż dane oryginalne, co jest oczywistą anomalią.

[1] https://archive.ics.uci.edu/ml/about.html.

[2] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data), Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Zgodnie z opisaną metodologią zastosowano proponowany algorytm do wybrania cech negatywnie wpływających na trafność. Tabela 1 przedstawia empiryczne *p*-wartości dla wybranych atrybutów[3].

Mając taki zestaw atrybutów, przystąpiono do konstruowania *n*-tych (zbiorów dwójek, trójek etc.), generując kombinacje w celu sprawdzenia, która z nich w największym stopniu poprawi trafność predykcji. Najlepszy rezultat testu *t*-Studenta różnicy trafności w 20-krotnej walidacji krzyżowej uzyskano, usuwając wszystkie wzmiankowane atrybuty.

Tabela 1. Selekcja cech nieistotnych

Atrybut	Empiryczna *p*-wartość
Inne osoby na utrzymaniu	≪0,001
Inne plany ratalne	≪0,001
Pracownik zza granicy	≪0,001
Stan cywilny i płeć	0,02
Liczba osób w stosunku do których zachodzi obowiązek alimentacyjny	0,03

Źródło: opracowanie własne.

Porównanie średniej trafności klasyfikatora bazowego oraz klasyfikatora po usunięciu atrybutów prezentuje tab. 2, wykres pudełkowy oraz wynik testu t-Studenta.

Tabela 2. Porównanie klasyfikatora bazowego i utworzonego przez usunięcie atrybutów

Miara\klasyfikator	Bazowy	Zredukowany
Średnia trafność zrównoważona (TZ)	0,5686	0,6131
Odchylenie standardowe TZ	0,038	0,05
1 kwantyl TZ	0,55	0,5923
Mediana TZ	0,566	0,6274
3 kwantyl TZ	0,5857	0,6518

Źródło: opracowanie własne.

[3] Zapewnienie powtarzalności wyników jest kluczowe dla rzetelności badań. Testy prowadzono, wykorzystując konkretne wersje bibliotek i narzędzi, a także z wykorzystaniem określonego sprzętu. Poniżej znajduje się specyfikacja wzmiankowanych elementów:

Komputer: MacBook Pro (Retina, 15-inch, Mid 2015), 2,2 GHz Intel Core i7, 16 GB 1600 MHz DDR3, Intel Iris Pro 1536 MB

System operacyjny: MacOS High Sierra 10.13.2 (17C205)

Środowisko analityczne: Rstudio 1.1.383

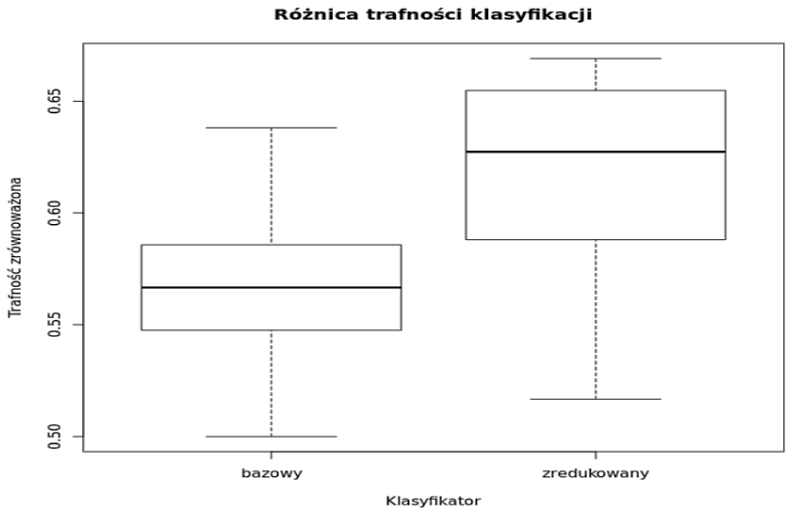
Wykorzystane biblioteki:

r-core: 3.4.2 – język analizy statystycznej R

RWeka 0.4-37 – budowanie klasyfikatora regułowego

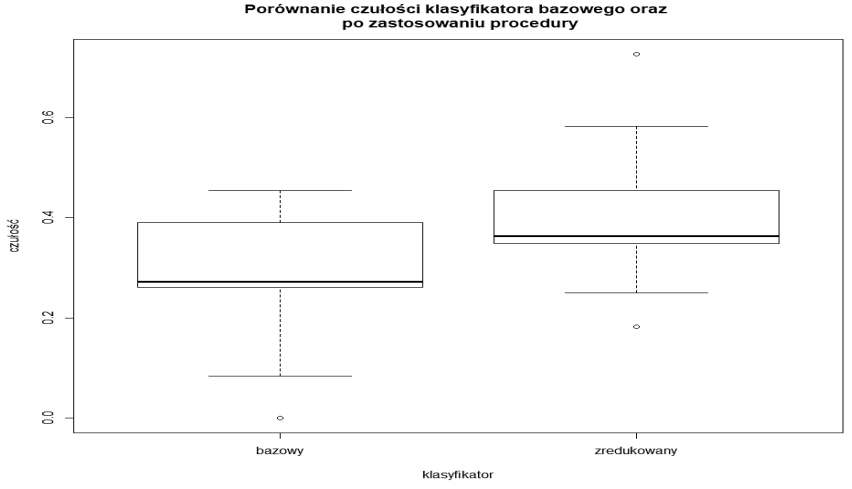
ststs 3.4.2. – testy statystycznej istotności

Różnica średnich TZ	-0,0445
Statystyka T	-2,2151
Stopnie swobody	18
P-wartość	0,019
Przedział 95% min	-∞
Przedział 95% max	-0,009



Rys. 1. Rozkład trafności klasyfikacji

Źródło: opracowanie własne.



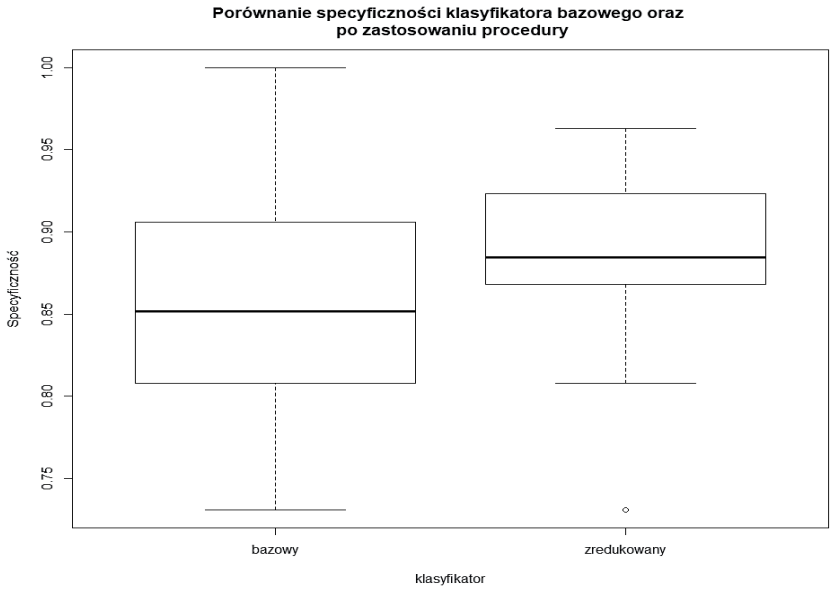
Rys. 2. Porównanie czułości klasyfikatorów

Źródło: opracowanie własne.

Warto również przeanalizować miary w rozbiciu na etykiety kategorii dla klasyfikatora korzystającego tylko z istotnych atrybutów. Uzyskał on następujące średnie rezultaty w 20-krotnej walidacji:

- **Specyficzność**: 88,7%, co jest wynikiem lepszym o prawie 3 punkty procentowe od klasyfikatora bazowego (85,9%),
- **Czułość**: 39%, co jest wynikiem lepszym o prawie 10 punktów procentowych od klasyfikatora bazowego (29%).

Rysunki 2 i 3 przedstawiają kolejno: relację specyficzności obu klasyfikatorów oraz porównanie czułości klasyfikatorów.



Rys. 3. Porównanie specyficzności klasyfikatorów

Źródło: opracowanie własne.

Analiza wyników pozwala stwierdzić, iż zachodzi statystycznie istotna różnica w trafności zrównoważonej na korzyść klasyfikatora RIPPER uzyskanego metodą permutacyjnego ustalania istotności atrybutów. Oznacza to, iż taki klasyfikator lepiej sprawdza się w podejmowaniu decyzji w przedmiotowym zbiorze danych. Odznacza się wyraźnie wyższą średnią czułością i specyficznością niż klasyfikator bazowy, co pozwala mu sprawniej identyfikować klientów.

6. Zakończenie

Jak wykazały badania empiryczne przeprowadzone w poprzednich sekcjach, zastosowanie permutacyjnej metody ustalania istotności atrybutów dla trafności zrównoważonej (BART) z algorytmem RIPPER pozwala otrzymać lepsze rezultaty w sto-

sunku do klasyfikacji bazowej. Uzyskany system odznacza się w stosunku do klasyfikatora bazowego:

1. Lepszym ogólnym wynikiem trafności zrównoważonej – nawet w warunkach silnego zaburzenia proporcji pomiędzy klasami.

2. Wyższą czułością i specyficznością.

Wymienione wyżej czynniki przekładają się na większą biznesową wartość finalnego systemu. Po pierwsze, wymaga on mniejszej ilości danych do działania, co w kontekście tzw. *Big Data* i obecności bardzo dużych wolumenów informacji ma niebagatelne znaczenie. Zidentyfikowanie atrybutów nieistotnych pozwoli z jednej strony zaoszczędzić zasoby obliczeniowe i czas przeprowadzanych operacji (korzyść w obszarze technologicznym), jak i potencjalnie uprościć procesy biznesowe zbierania danych. Po drugie, wyższa czułość i specyficzność przekładają się na system stabilniejszy, zachowujący się w sposób przewidywalny dla użytkownika, a tym samym wzbudzający większe zaufanie. Peter Flach wskazuje to jako jeden z istotnych biznesowo czynników kształtujących dobrą architekturę wspomagania decyzji biznesowych [Flach 2012, s. 93].

Wnioski przedstawione w niniejszym artykule nie wyczerpują tematu selekcji atrybutów istotnych dla algorytmów regułowych. Poruszone w tym miejscu zagadnienia stanowią przyczynek do dalszych badań. Wśród najistotniejszych, zdaniem autorów, znajdują się:

1. Optymalizacja opisanego algorytmu pod kątem złożoności cyklomatycznej i efektywności – przedstawiona implementacja jest tzw. wdrożeniem „naiwnym", a co za tym idzie – niezoptymalizowanym. Może to rodzić problemy w przypadku dużych wolumenów danych i należy zbadać możliwości przyspieszenia procesu.

2. Zbadanie skuteczności metody na większej liczby zbiorów danych, zróżnicowanych pod kątem proporcji klas.

3. Sprawdzenie, czy istnieje związek pomiędzy statystycznymi miarami dopasowania reguł, a ogólną trafnością predykcyjną – możliwe byłoby wówczas wyprowadzenie rozwiązania analitycznego.

Literatura

Aggarwal C.C., 2015, *Data classification : algorithms and applications*, Boca Raton, CRC.

Breiman L., 2001, *Random forests*, Machine Learning, 45(1), s. 5-32.

Brodersen K.H. i in., 2010, *The Balanced Accuracy and Its Posterior Distribution*, [w:] *Proceedings*, IEEE Computer Society Press, Washington D.C., s. 3121-3124.

Chomiak-Orsa I., Mrozek B., 2017, *Analiza wielkich zbiorów danych w mediach społecznościowych – perspektywa przedsiębiorcy,* Przegląd Organizacji, 8, s. 48-55.

Cohen W.W., 1995, *Fast Effective Rule Induction*, [w:] Proc. 12th International Conference on Machine Learning.

Eibe F., Witten I.H., 1998, *Using a Permutation Test for Attribute Selection in Decision Trees*, [w:] *International Conference on Machine Learning*, s. 152-160.

Flach P.A., 2012, *Machine Learning: the Art and Science of Algorithms That Make Sense of Data*, Cambridge: Cambridge University Press.

Good P., 1994, *Permutation Tests:A Practical Guide to Resampling Methods for Testing Hypotheses*, NY: Springer New York (Springer Series in Statistics), New York.

Hastie T., Tibshirani R., Friedman J., 2009, *The Elements of Statistical Learning*., 2, Springer.

Kisielecki J., 2008, *Systemy informatyczne zarządzania*, Placet, Warszawa.

Mitchell T., 1997, *Machine Learning*, McGraw-Hill, New York.

Morzy T., 2013, *Eksploracja danych: metody i algorytmy*, Wydawnictwo Naukowe PWN, Warszawa.

Ojala M., Garriga G.C., 2010, *Permutation tests for studying classier performance*, Journal of Machine Learning Research, 11, s. 1833-1863.

Olszak C.M., Mach-Król M., 2016, *Big Data: How to Gain Value for Organizations*, European, Mediterranean & Middle Eastern Conference on Information Systems 2016 (EMCIS2016), June 23th-24th 2016, Krakow, Poland.

Radivojac P. i in., 2004, *Feature Selection Filters Based on the Permutation Test,* [w:] Boulicaut J.-F. i in. (red.) *Machine Learning: ECML 2004*, Heidelberg: Springer Berlin Heidelberg, Berlin, s. 334-346.

Radivojac P. i in., 2004, *Feature Selection Filters Based on the Permutation Test*, [w:] Boulicaut J.-F. i in. (red.), *Machine Learning: ECML 2004*, Heidelberg: Springer Berlin Heidelberg, Berlin, s. 334-346.

Surma J., 2009, *Business Intelligence : systemy wspomagania decyzji biznesowych*, wyd. 1, 2, Wydawnictwo Naukowe PWN, Warszawa.