ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2015
Seria: ORGANIZACJA I ZARZĄDZANIE z. 86 Nr kol. 1946

Tomasz OWCZAREK, Adam SOJDA
Politechnika Śląska
Wydział Organizacji i Zarządzania
tomasz.owczarek@polsl.pl; adam.sojda@polsl.pl
Konrad KACZMAREK
Politechnika Śląska
Wydział Matematyki Stosowanej
konrad.kaczmarek@polsl.pl

WPŁYW LICZBY PREDYKTORÓW NA SKUTECZNOŚĆ ALGORYTMÓW OPARTYCH NA DRZEWACH KLASYFIKACYJNYCH

Streszczenie. Współczesne organizacje, aby być konkurencyjne, muszą mieć umiejętności przetworzenia olbrzymich danych. Jednym z najbardziej obiecujących kierunków w tym zakresie jest wykorzystanie analityki predykcyjnej, opierającej się na algorytmach i modelach uczenia maszynowego. Związanych z tym jest wciąż wiele wyzwań, m.in. pytanie o „wejście" do takich modeli, czy powinny to być wszystkie dane zgromadzone przez organizację czy może raczej wcześniej wybrane zmienne? Celem artykułu jest zbadanie skuteczności algorytmów opartych na drzewach klasyfikacyjnych ze względu na liczebność predyktorów.

Słowa kluczowe: klasyfikacja, dobór zmiennych, drzewa klasyfikacyjne, analityka predykcyjna.

THE INFLUENCE OF NUMBER OF PREDICTORS ON ACCURACY OF CLASSIFICATION ALGORITHMS BASED ON TREES

Summary. To stay competitive contemporary organizations have to master in processing massive amount of data. Predictive analytics, that is analytics based on machine learning algorithms and models, is one of the most promising directions. But there are many issues involved. One of them is the input to such models: should it be all data gathered by organization or just the selected variables? The aim of the article is to check how the number of predictors influences accuracy of classification algorithms based on trees.

Keywords: classification, feature selection, classification trees, predictive analytics.

1. Wprowadzenie

Jednym z podstawowych wyzwań, przed którym stają współczesne przedsiębiorstwa, jest poradzenie sobie z coraz większą ilością danych – pochodzących zarówno z wnętrza przedsiębiorstw, jak i z ich otoczenia. Nieustannie rosnący wolumen danych wynika przede wszystkim z rejestracji cyfrowych śladów działalności człowieka na skutek powszechnego wykorzystywania sieci Internet oraz urządzeń mobilnych (wraz z rozwojem takich zjawisk jak Internet rzeczy czy media społecznościowe), ale także z postępu technologicznego, dzięki któremu możliwe jest gromadzenie coraz bardziej szczegółowych danych z coraz większą częstotliwością [15]. Swobodny dostęp do olbrzymiej ilości danych powoduje, że przewagę konkurencyjną uzyskają te przedsiębiorstwa, które będą w stanie je najlepiej wykorzystać [13, s. 315-321], czego dowodzą m.in. badania E. Brynjolfssona i L.M. Hitta [3].

Efektem nastawienia przedsiębiorstw na możliwie najpełniejsze wykorzystanie dostępnych danych jest wykształcenie się nowego podziału w analityce biznesowej – na tradycyjnie pojmowane *business intelligence* oraz na tzw. zaawansowaną analitykę (*advanced analytics*). Do tej pierwszej zalicza się analitykę deskryptywną oraz (częściowo) diagnostyczną, których celem jest opisanie aktualnego stanu. Zaawansowana analityka natomiast zawiera w sobie analityki predykcyjną oraz normatywną (preskryptywną), których zadaniem jest udzielenie odpowiedzi na pytania o możliwe przyszłe stany [10].

W zaawansowanej analityce równie dużą rolę, co tradycyjne metody statystyczne, o ile nawet nie większą, odgrywają algorytmy uczenia maszynowego (*machine learning*), takie jak drzewa klasyfikacyjne, maszyny wektorów wspierających (*support vector machines*) czy sieci neuronowe. W algorytmach tych, w odróżnieniu od analitycznych modeli ekonometrycznych, nie zakłada się niczego o teoretycznym mechanizmie generującym dane – nie przeprowadza się tutaj testów zgodności rozkładu czy współliniowości. Jedynym kryterium oceny ich adekwatności jest ich skuteczność predykcyjna, mierzona na specjalnie wydzielonym zbiorze testowym, który nie jest wykorzystany do oszacowania parametrów tych modeli [2]. Pojawia się wobec tego pytanie o to, co ma stanowić wejście dla tych modeli: czy lepiej, jeśli przedsiębiorstwa wykorzystują wszystkie zgromadzone przez siebie dane czy też jakość predykcji istotnie wzrośnie, jeśli przed zastosowaniem odpowiednich algorytmów dokona się starannego wyboru zmiennych, które posłużą jako predyktory? Pytanie to nie ma jednoznacznej odpowiedzi, a oba podejścia niosą ze sobą pewne ryzyko. Pierwsze wiąże się z tzw. przekleństwem wielowymiarowości, przejawiającym się w tym, że wzrost liczby wymiarów (tzn. predyktorów), przy stałej liczebności próbki służącej do oszacowania modelu, może spowodować spadek jego skuteczności predykcyjnej [9, s. 22-27]. Drugie podejście, tzn. wcześniejsza selekcja zmiennych, wobec braku metod, które w sposób jednoznaczny wskażą wszystkie istotne zmienne [7], grozi wykluczeniem predyktorów, które mogą wnosić wartościową informację o wariancji zmiennej predykowanej.

Celem niniejszego artykułu jest przetestowanie, w jaki sposób wzrost liczby zmiennych służących do predykcji wpłynie na skuteczność algorytmów klasyfikacyjnych opartych na drzewach decyzyjnych. Wybór właśnie tych algorytmów podyktowany jest ich odpornością na anomalie i zróżnicowanie zmiennych znajdujących się na ich wejściu, co powoduje, że są one szczególnie użyteczne w rzeczywistych przypadkach biznesowych [11, s. 174].

Dalszą część artykułu zorganizowano w następujący sposób. W punkcie drugim zamieszczono założenia leżące u podstaw przeprowadzonych badań oraz ich szczegółowy opis. W punkcie trzecim zaprezentowano wyniki symulacji. W punkcie czwartym znajdują się wnioski z badań, ich implikacje oraz potencjalne kierunki dalszych prac.

2. Opis przeprowadzonych badań

2.1. Przygotowanie danych

Z punktu widzenia celu badań, szczególnie istotne były dwie kwestie:

1) testowane przypadki powinny w miarę możliwości oddawać swoim charakterem rzeczywiste problemy, z którymi mierzy się przedsiębiorstwo,
2) powinna istnieć możliwość porównania skuteczności wybranych algorytmów dla zmiennej liczby predyktorów.

Ze względu na pierwsze założenie zdecydowano, że predykowana zmienna *Y* będzie zmienną kategoryczną o trzech możliwych wartościach (A, B i C) i nierównomiernym rozkładzie[1], natomiast zmienne stanowiące predyktory powinny różnić się pod względem rozkładu i charakteru (miały się wśród nich znaleźć zmienne ciągłe, całkowite i kategoryczne). Drugie założenie wymagało, żeby skuteczność identyfikacji wartości zmiennej *Y* na podstawie wartości predyktorów mogła być z góry oszacowana. Ostatecznie zdecydowano się na 4 zmienne (oznaczane dalej jako *X1*, *X2*, *X3* i *X4*), które miały pozwolić na pełną identyfikację zmiennej *Y* wg ustalonych z góry reguł. Wartości tych zmiennych zostały wygenerowane losowo (dla każdej zmiennej wygenerowano 10 tys. wartości) na podstawie rozkładów przedstawionych w tabeli 1.

Jeśli chodzi o zmienną *Y*, to reguły służące do jej wygenerowania ilustruje rys. 1. Wartości dwóch pierwszych predyktorów (czyli *X1* i *X2*, przedstawione na rys. 1 odpowiednio na osiach X i Y) pozwalały oddzielić kategorię A zmiennej *Y* (kolor czerwony) od kategorii B i C na podstawie kwartyli zmiennych *X1* i *X2*. Przykładowo, jeżeli wartość

[1] Prostszym rozwiązaniem byłoby przyjęcie jako *Y* zmiennej binarnej. Wydaje się jednak, że większa liczba kategorii lepiej odpowiada rzeczywistym problemom, z którymi mierzą się przedsiębiorstwa. Przykładowo, w konkursie sponsorowanym przez firmę Otto Group w serwisie Kaggle uczestnicy mieli za zadanie zaklasyfikować produkty do jednej z 9 kategorii (https://www.kaggle.com/c/otto-group-product-classification-challenge/data, dostęp: 29.09.2015).

zmiennej *X1* znajdowała się poniżej pierwszego kwartyla tej zmiennej oraz wartość zmiennej *X2* była poniżej pierwszego kwartyla tej zmiennej, to (bez względu na wartości pozostałych zmiennych) zmienna *Y* przyjmowała wartość A (w każdym z ośmiu paneli przedstawionych na rys. 1 sytuację tę obrazuje skrajny lewy dolny róg wykresu – znajdują się w nim jedynie czerwone punkty). Zmienne *X3* oraz *X4* pozwalały na rozróżnienie między kategoriami B i C zmiennej Y. I tak na przykład, jeśli zmienna *X3* przyjmowała wartości równe J1 lub J2 oraz zmienna *X4* była poniżej swojej mediany (na rys. 1 sytuację tę reprezentują dwa dolne panele z lewej strony), to zmienna *Y* mogła przyjąć jedynie wartości A lub B (zgodnie z tym, co napisano wcześniej, dalsze rozróżnienie między kategoriami A i B mogło być dokonane na podstawie wartości *X1* i *X2*). Przyjęte reguły budują „prostokątne" obszary podziału wartości zmiennej *Y*, z których rozpoznaniem dobrze radzą sobie algorytmy oparte na drzewach decyzyjnych [9, s. 305].

Tabela 1

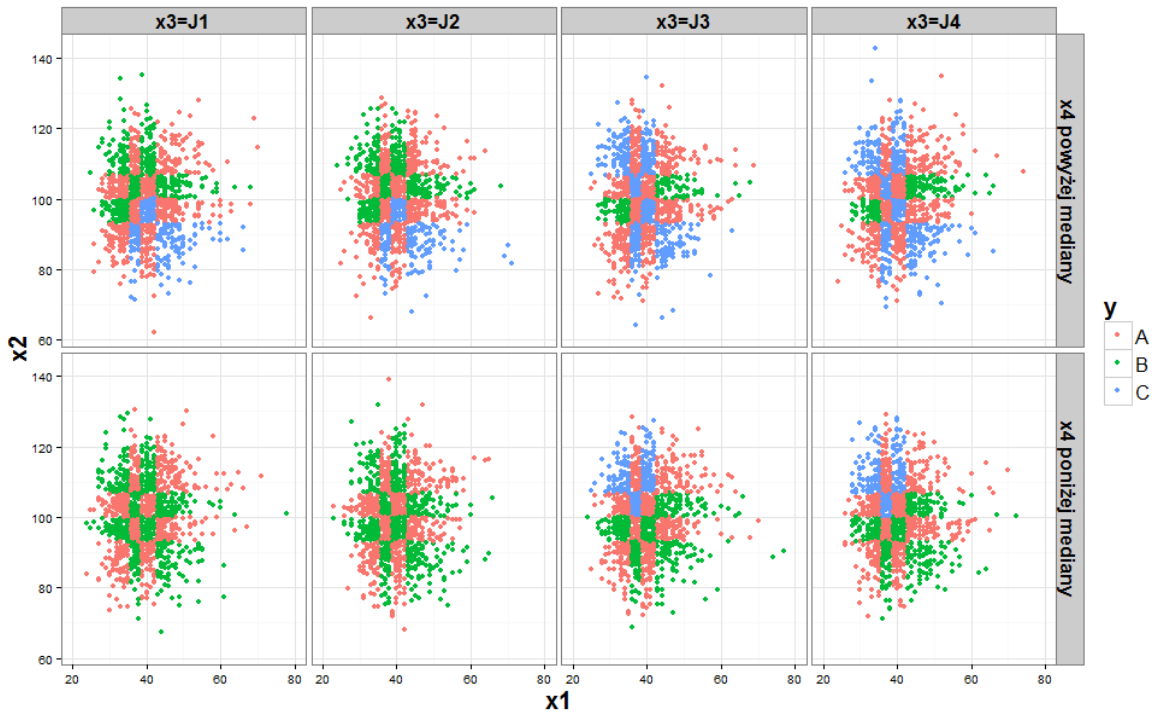
Opis podstawowego zbioru danych wykorzystanego w symulacjach

Zmienna	Opis zmiennej
X1	zmienna całkowita, rozkład normalny (średnia=35, odch. stand.=4) + rozkład wykładniczy (λ=0,2), zaokrąglona do liczby całkowitej
X2	zmienna ciągła, rozkład normalny (średnia 100, odchylenie 10)
X3	zmienna kategoryczna z czterema wartościami: J1, J2, J3, J4 rozłożonymi równomiernie
X4	zmienna ciągła, rozkład log-normalny (średnia=8, odchylenie standardowe=0,8)
Y	zmienna kategoryczna z trzema wartościami: A, B i C, częstość wartości: A=49,9%, B=31,5%, C=18,6%

Źródło: opracowanie własne.

Podane na początku tego punktu założenia wpłynęły na ostateczny kształt zbiorów danych, które posłużyły do eksperymentów symulacyjnych. Ponieważ celem badań było sprawdzenie wpływu liczebności predyktorów na skuteczność klasyfikacji, więc zdecydowano się przeprowadzić dwa warianty eksperymentów, w których do podstawowego zbioru danych dołożono odpowiednio 10 oraz 100 dodatkowych zmiennych predykcyjnych. Należało przy tym uwzględnić fakt, że rzeczywiste dane, którymi dysponują przedsiębiorstwa, mogą być w różnym stopniu powiązane ze zmienną predykowaną. To oznaczało, że pomimo istnienia czterech zmiennych, które w pełni pozwalały na rozpoznanie wartości zmiennej *Y*, wśród pozostałych zmiennych również mogły się znajdować wartościowe predyktory. Z tego względu w każdym z wymienionych wariantów rozpatrywano 4 przypadki. Przypadek 0 oznaczał, że dodatkowe predyktory nie są powiązane ze zmienną *Y* – ich wartości były wygenerowane w sposób całkowicie losowy. Przypadki 1-3 ilustrowały sytuacje, w których dodatkowe predyktory powiązane były ze zmienną *Y*

w różnym stopniu (najsłabiej w przypadku 1, najmocniej w przypadku 3). Szczegółowy opis poszczególnych przypadków znajduje się w tabeli 2.



Rys. 1. Rozkład wartości zmiennej *Y* ze względu na wartości zmiennych *X1*, *X2*, *X3* i *X4*
Fig. 1. Distribution of variable *Y* with respect to *X1*, *X2*, *X3* and *X4*
Źródło: opracowanie własne.

Tabela 2

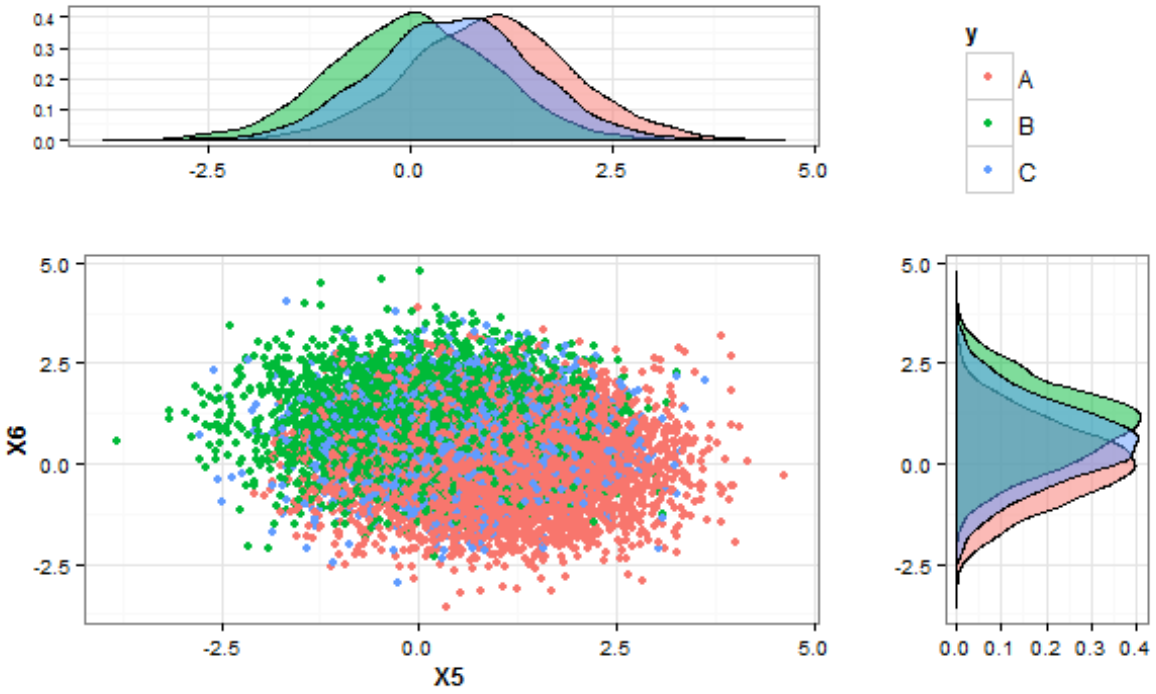
Opis przypadków dodatkowych zmiennych predykcyjnych

Przypadek	Opis
0	wartości dodatkowych zmiennych mają rozkład normalny ze średnią równą 0 i odchylenie standardowe równym 1
1	wartości dodatkowych zmiennych mają rozkład normalny z odch. stand. równym 1 i średnią równą 0, 0,1 lub 0,2 w zależności od wartości zmiennej *Y*
2	wartości dodatkowych zmiennych mają rozkład normalny z odch. stand. równym 1 i średnią równą 0, 0,3 lub 0,6 w zależności od wartości zmiennej *Y*
3	wartości dodatkowych zmiennych mają rozkład normalny z odch. stand. równym 1 i średnią równą 0, 0,5 lub 1 w zależności od wartości zmiennej *Y*

Źródło: opracowanie własne.

Przykładowy związek między dodatkowymi zmiennymi predykcyjnymi a zmienną *Y* zilustrowano na rys. 2. Przedstawia on wartości dwóch dodatkowych zmiennych (oznaczonych jako *X5* i *X6*) z wariantu nr 3. Górny oraz prawy wykres ilustrują rozkłady tych dwóch zmiennych w podziale ze względu na wartości *Y*. Można z nich odczytać, że gdy *Y* przyjmuje wartość (kategorię) B (kolor zielony), to w tym przypadku średnia rozkładu

zmiennej *X5* wynosi 0, a średnia rozkładu zmiennej *X6* znajduje się w okolicach 1. Pozwala to na częściową predykcje zmiennej *Y* na podstawie zmiennych *X5* i *X6*. Przykładowo, dla niskich wartości zmiennej *X5* i wysokich wartości zmiennej *X6* (lewa górna ćwiartka środkowego wykresu) istnieje duże prawdopodobieństwo, że zmienna *Y* przyjmie wartość B.



Rys. 2. Rozkład wartości zmiennej *Y* ze względu na wartości zmiennych *X5* i *X6* (przypadek nr 3)
Fig. 2. Distribution of variable *Y* with respect to *X5* nad *X6* (case no. 3)
Źródło: opracowanie własne.

Ostatecznie przygotowano 9 zbiorów danych, każdy liczący po 10 tys. rekordów:

- *d0* – zbiór składający się ze zmiennej *Y* oraz zmiennych *X1*, *X2*, *X3* i *X4*,
- cztery zbiory *d10.x* (*x*=1,…4) – zbiór d0 z dołączonymi 10 dodatkowymi zmiennymi,
- cztery zbiory *d100.x* (*x*=1,…4) – zbiór d0 z dołączonymi 100 dodatkowymi zmiennymi[2].

2.2. Procedura eksperymentów symulacyjnych

Do testowania wybrano następujące algorytmy: drzewo CART [9, s. 305-310], pojedynczy model C5.0 [11, s. 392-396], *random forest* [1], *boosted C5.0* [11, s. 396-397], adaBoost.M1 [4] oraz *stochastic gradient boosting* [5]. Oprócz przyjętego celu badań interesujące wydawały się ewentualne różnice między pojedynczymi modelami (drzewo CART i model C5.0) a modelami „złożonymi", tzn. składającymi się z większej liczby modeli predykcyjnych (pozostałe 4 algorytmy), a także różnice między modelem *random forest*

[2] Symbol *x* w oznaczeniu zbioru danych to numer przypadku, przykładowo d10.0 określa zbiór, w którym dodatkowe 10 zmiennych ma rozkład w pełni losowy, niepowiązany ze zmienną *Y*.

(w którym poszczególne drzewa predykcyjne wybierane są w sposób losowy) a algorytmami opartymi o procedurę *boostingu* (w których predyktory generowane są w kolejnych iteracjach z uwzględnieniem ujemnych wag za źle sklasyfikowane przypadki [6]). Ze względu na fakt, że algorytmy te posiadają różną liczbę parametrów dostrojenia, zdecydowano się przyjąć następujący przebieg procedury testowej:

1) podział zbioru danych na dwa podzbiory: treningowy i testowy (w stosunku 70%-30%),
2) dobór parametrów każdego modelu na podstawie danych ze zbioru treningowego, stosując 10-krotną kroswalidację (*10-fold cross validation*),
3) ocena skuteczności najlepszego wariantu danego modelu (wyznaczonego na podstawie kroswalidacji) na zbiorze testowym.

Do testowania wykorzystano implementację wymienionych algorytmów w pakietach języka R[3]. Parametry poszczególnych modeli, które były testowane w trakcie kroswalidacji, przedstawiono w tabeli 3.

Tabela 3

Testowane parametry modeli w trakcie kroswalidacji

Przypadek	Opis
CART	parametr złożoności c_p
C5.0	rodzaj modelu (drzewo albo reguły)
random forest	m_{try} (liczba zmiennych wybieranych przy każdym podziale)
boosted C5.0	rodzaj modelu, liczba iteracji
adaBoost	maksymalny rozmiar drzewa, liczba iteracji
grBoosting	maksymalny rozmiar drzewa, liczba iteracji, parametr η

Źródło: opracowanie własne.

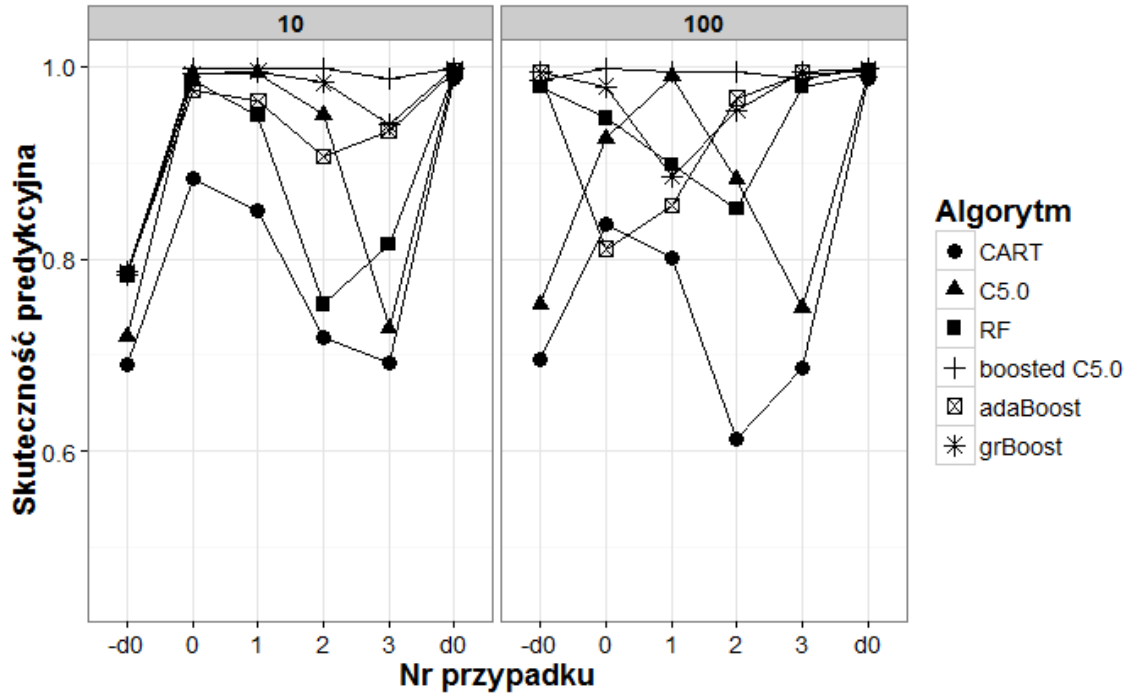
3. Wyniki przeprowadzonych symulacji

Rezultaty przeprowadzonych badań przedstawiono na rys. 3. Składa się on z dwóch paneli: lewy zawiera wyniki dla 10 dodatkowych zmiennych, a prawy dla 100. Na obu panelach, oprócz wyników dla opisywanych wcześniej 4 przypadków, przedstawiono również rezultaty dodatkowych wariantów przeprowadzonych symulacji: **d0** opisuje wariant, w którym do predykcji zmiennej *Y* wykorzystano jedynie zmienne *X1*, *X2*, *X3* i *X4*[4], natomiast **–d0** to warianty, w których do predykcji wykorzystano jedynie dodatkowe zmienne z przypadku 3 (odpowiednio 10 i 100 dodatkowych zmiennych) – tzn. te, w których dodatkowe zmienne były najsilniej powiązane ze zmienną *Y*, ale bez zmiennych *X1*, *X2*, *X3* i *X4*. Warianty te dołożono, aby móc lepiej porównać wpływ dodatkowych predyktorów na skuteczność predykcyjną. Na osi Y przedstawiono skuteczność predykcyjną algorytmów

[3] Wykorzystano do tego pakiety „rpart", „C50", „randomForest", „adabag" , „caret" oraz „xgboost".
[4] W obu panelach znajdują się te same wartości dla tego wariantu.

mierzoną jako stosunek liczby poprawnie sklasyfikowanych wartości zmiennej *Y* do wszystkich liczby wszystkich danych w zbiorze testowym.

W pierwszej kolejności należy zaznaczyć, że jeśli jako predyktory wykorzystane były jedynie zmienne *X1*, *X2*, *X3* i *X4* (ostatnia kolumna na obu panelach) wszystkie testowane algorytmy bardzo dobrze poradziły sobie z przygotowanym zbiorem danych. Najniższą skuteczność (98,7%) uzyskało tutaj drzewo CART, wyniki pozostałych modeli wyniosły powyżej 99%.



Rys. 3. Rezultaty przeprowadzonych symulacji
Fig. 3. Results of conducted simulations
Źródło: opracowanie własne.

Porównanie przypadków 0 i 1 w poszczególnych wariantach pozwala udzielić odpowiedzi na postawione pytanie badawcze. Ze wszystkich algorytmów jedynie *boosted C5.0* okazał się odporny na obecność większej liczby dodatkowych predyktorów. Dla pozostałych algorytmów obserwuje się wyraźny spadek ich skuteczności w wariancie ze 100 dodatkowymi zmiennymi. W szczególności, jak można odczytać z wykresu, dołożenie dodatkowych 10 predyktorów niezwiązanych ze zmienną predykowaną (przypadek nr 0 na lewym panelu) w wyraźny sposób wpłynęło jedynie na skuteczność predykcji pojedynczego drzewa CART. Wprowadzenie 100 takich zmiennych (przypadek nr 0, prawy panel) spowodował obniżenie skuteczności wszystkich algorytmów (za wyjątkiem wspomnianego wcześniej *boosted C5.0*).

Wyniki w obu wariantach dla przypadków 0, 1 i 2 pokazują, że wzrost zależności między zmienna predykowaną a dodatkowymi predyktorami powodują w większości algorytmów spadek ich skuteczności predykcyjnej. Wyjątkami od tej reguły są pojedynczy model C5.0

(który dla 100 zmiennych w przypadku 1 uzyskał wyniki zbliżone do swojej bardziej złożonej wersji) oraz algorytm *adaBoost.M1* (który w wariancie dla 100 dodatkowych zmiennych wraz z rosnącym numerem przypadku poprawiał swoje rezultaty). Generalnie należy zauważyć, że w testowanych przypadkach algorytmy oparte na procedurze *boostingu* radzą sobie lepiej z dodatkowymi zmiennymi niż pojedyncze modele czy algorytm *random forest*.

Interesujące rezultaty zaobserwowano dla przypadku nr 3. Po pierwsze, wyższy stopień powiązania dodatkowych predyktorów ze zmienną *Y* spowodował nieznaczne obniżenie skuteczności algorytmu *boosted C5.0*. Dla pozostałych „złożonych" modeli przyczynił się jednak do wzrostu ich skuteczności (z jednym wyjątkiem – algorytmu *stochastic gradient boosting* przy 10 dodatkowych zmiennych). W szczególności warto zwrócić uwagę na wyniki z przypadku 3 oraz –d0 dla 100 dodatkowych zmiennych, które są bardzo zbliżone, a dla modeli „złożonych" nie odbiegają w sposób wyraźny od wyników z przypadku d0. Z jednej strony świadczy to o tym, że obecność dużej liczby predyktorów powiązanych ze zmienną predykowaną może „przesłonić" te zmienne, które pozwoliłyby na jej pełną identyfikację. Z drugiej jednak, jak wskazują uzyskane wyniki, modele „złożone" są w stanie uwzględnić w swoich predykcjach informacje wnoszone przez gorsze predyktory, co w konsekwencji podnosi skuteczność ich predykcji.

4. Wnioski

Uzyskane rezultaty wskazują, że z obecnością dodatkowych zmiennych predykcyjnych najlepiej radzi sobie algorytm *boosted C5.0*. Należy jednak zwrócić uwagę na fakt, że przeprowadzone eksperymenty miały charakter laboratoryjny, a testowane zbiory danych były wygenerowane w sposób sztuczny. W rzeczywistych przypadkach raczej trudno spodziewać się występowania takich zmiennych, które pozwoliłyby na pełną identyfikację wartości zmiennej predykowanej.

Podstawowy wniosek, który nasuwa się na podstawie przeprowadzonych badań jest taki, że algorytmy oparte na procedurze *boostingu* są bardziej odporne na obecność dodatkowych zmiennych zakłócających predykcję. Jednak nawet one nie zawsze są zdolne do uwzględnienia w wygenerowanych modelach jedynie tych informacji, które są niezbędne do identyfikacji wartości zmiennej predykowanej. Dlatego też nie należy zaniedbywać procedury wyboru właściwych predyktorów, bo może ona mieć decydujący wpływ na skuteczność opracowanych modeli. Dodatkowo, co pokazały przeprowadzone eksperymenty, możliwe jest uzyskanie (dzięki temu) prostszych, a tym samym łatwiejszych do interpretacji modeli [8], których jakość predykcji będzie porównywalna do predykcji uzyskanych za pomocą modeli bardziej złożonych. Ma to szczególne znaczenie w kontekście biznesowym [14, s. 117].

Oczywiście, w dalszym ciągu aktualny pozostaje problem wspomniany na początku niniejszego artykułu – tzn. ryzyko utraty informacji na skutek odrzucenia wartościowych predyktorów. Należy w tym miejscu zauważyć, że testowane algorytmy dają możliwość oceny zmiennych predykcyjnych pod względem ich ważności dla zmiennej predykowanej. Istnieją również algorytmy wskazujące istotne zmienne, które wykorzystują tę ich charakterystykę [12]. Połączenie wskazań modeli oraz tradycyjnych miar statystycznej zależności między zmiennymi powinno zmniejszyć ryzyko odrzucenia wartościowych predyktorów.

Problemem, który nie został poruszony w niniejszym opracowaniu, jest złożoność obliczeniowa oraz czas potrzebny na wytrenowanie modeli, który zwiększa się wraz ze wzrostem liczby zmiennych. Właśnie ze względu na czasochłonność procedury kroswalidacji, która służyła do optymalnego doboru parametrów modeli (co wymagało niekiedy wygenerowania kilkuset realizacji modelu dla danego algorytmu), w każdym z opisywanych wariantów skuteczność modeli testowana była jedynie na jednym zbiorze danych. Dla uzyskania bardziej pewnych wniosków na temat skuteczności testowanych algorytmów w określonych przez warunki eksperymentu wariantach, należałoby oszacować wpływ czynnika losowego, związanego z testowanymi danymi. Wymagałoby to przynajmniej kilkukrotnego powtórzenia każdego testowanego przypadku na różnych zbiorach danych.

Bibliografia

1. Breiman, L.: Random forests, Machine learning, Vol. 45, No. 1, 2001, p. 5-32.
2. Breiman L.: Statistical Modeling: The Two Cultures, Statistical Science, Vol. 16, No. 3, 2001, p. 199-231.
3. Brynjolfsson E., Lorin M.H., Kim H.H.: Strength in numbers: How does data-driven decisionmaking affect firm performance?, SSRN Electronic Journal, April 2011 (http://dx.doi.org/10.2139/ssrn.1819486).
4. Freund, Y., Schapire R.E.: Experiments with a new boosting algorithm, Proceedings of the Thirteenth International Conference on Machine Learning, 1996, p. 148-156.
5. Friedman, J.H.: Stochastic gradient boosting, Computational Statistics & Data Analysis Vol. 38, No. 4, 2002, p. 367-378.
6. Friedman J., Hastie T., Tibshirani R.: Additive Logistic Regression: A Statistical View of Boosting, The Annals of Statistics, Vol. 28, No. 2, p. 337-374.
7. Guyon I., Elisseeff A.: An Introdustion to Variable and Feature Selection, Journal of Machine Learning Research 3, 2003, p. 1157-1182.
8. Hand, D.J.: Classifier technology and the illusion of progress, Statistical science, Vol. 21, No. 1, 2006, p. 1-15.

9. Hastie T., Tibshirani R., Friedman J.: The Elements of Statistical Learning, Springer, New York 2009.
10. Koch R.: From Business Intelligence to Predictive Analytics, Strategic Finance 96(7), 2015, p. 56-57.
11. Kuhn M., Johnson K.: Applied Predictive Modeling, Springer, New York, 2013.
12. Kursa M.B., Rudnicki W.R.: Feature Selection with Boruta Package, Journal of Statistical Software, Vol. 36, No. 11, 2010
13. Provost F., Fawcett T.: Data Science for Business, O'Reilly Media, 2013.
14. Schutt R., O'Neil C., Doing Data Science. Straight Talk from the Frontline, O'Reilly Media, 2014.
15. Wielki J.: Implementation of the Big Data concept in organizations – possibilities, impediments and challenges, Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, 2013, p. 985–989.

Abstract

To stay competitive contemporary organizations have to master in processing massive amount of data. Predictive analytics, that is analytics based on machine learning algorithms and models, is one of the most promising directions. But there are many issues involved. One of them is the input to such models: should it be all data gathered by organization or just the selected variables?

The aim of the article was to examine how the number of predictors influences accuracy of classification algorithms based on trees. In order to answer this question several simulation experiments were conducted in which the accuracy of six different classification algorithms were measured. The results are presented in Fig. 3. The main conclusion is that larger number of unnecessary predictors has negative impact on the accuracy, although the effect is smaller for algorithms based on boosting procedure.