Raport Badawczy
Research Report

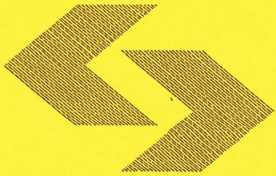
RB/19/2014

Odporność regresyjnych metod klasyfikacji binarnej na odstępstwa od podstawowych założeń

O. Hryniewicz, J. Karpiński,
A. Olwert

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences



POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 3810100

fax: (+48) (22) 3810105

Kierownik Zakładu zgłaszający pracę:
Prof. dr hab. inż. Olgierd Hryniewicz

Warszawa 2014

OLGIERD HRYNIEWICZ
Wyższa Szkoła Informatyki Stosowanej i Zarządzania w Warszawie
JANUSZ KARPIŃSKI
Instytut Badań Systemowych PAN
ANNA OLWERT
Instytut Badań Systemowych PAN

ODPORNOŚĆ REGRESYJNYCH METOD KLASYFIKACJI BINARNEJ NA ODSTĘPSTWA OD PODSTAWOWYCH ZAŁOŻEŃ

Streszczenie

W pracy przedstawiono analizę przydatności klasyfikatorów binarnych, wykorzystujących równania prostej regresji liniowej oraz regresji kwadratowej powierzchni odpowiedzi wykorzystywanych do realizacji zadań klasyfikacji, a także regresji logistycznej, w przypadku złożonych i nieliniowych związków pomiędzy binarną zmienną klasyfikującą i zmiennymi objaśniającymi opisanymi rozkładami różnymi od rozkładu normalnego. W analizie uwzględniono również przypadki, gdy zbiory uczące, na podstawie których konstruowano klasyfikatory, istotnie różnią się od zbiorów testujących (lub danych spotykanych w zastosowaniach). Przeprowadzony eksperyment symulacyjny sugeruje, że w przypadku tego samego modelu opisującego zbiory uczące i zbiory testujące zastosowanie klasyfikatora opartego na kwadratowej powierzchni odpowiedzi daje lepsze rezultaty niż w przypadku zastosowania prostej regresji liniowej. Z kolei, klasyfikatory oparte na prostej binarnej regresji liniowej są bardziej odporne na zmianę modelu danych testujących.

Słowa kluczowe: klasyfikatory binarne, regresja liniowa, regresja kwadratowej powierzchni odpowiedzi, regresja logistyczna, kopuły

1. Wprowadzenie

Zadania klasyfikacji są podstawowym przedmiotem badań w zakresie eksploracji danych (*data mining*). Można je przedstawić jako metody analizy *wielowymiarowych* cech statystycznych, z których przynajmniej jedna jest cechą dyskretną. Zadanie klasyfikacji polega w tym przypadku na przewidzeniu wartości tej wyróżnionej cechy dyskretnej (zmiennej klasyfikującej) na podstawie znajomości wartości pozostałych cech statystycznych. Jeżeli interesująca nas dyskretna cecha statystyczna przyjmuje tylko dwie wartości, to w takim przypadku mówimy o klasyfikacji binarnej. W ogólnym przypadku wartościami takiej cechy mogą być dowolne *etykiety*. Jednakże w zadaniach, które będziemy rozpatrywać w niniejszej pracy będziemy przyjmować, że interesująca nas cecha będzie przyjmować dwie możliwe wartości liczbowe, zazwyczaj *zero* lub *jeden*. Funkcje wartości zmiennych

opisujących klasyfikowane obiekty, które przypisują zmiennej klasyfikującej konkretne wartości nazywamy *klasyfikatorami*. W szczególnym przypadku, gdy możliwe są tylko dwie wartości zmiennej klasyfikującej, możemy mówić o *klasyfikatorach binarnych*.

Zadania klasyfikacji po raz pierwszy zostały sformułowane w obszarze statystyki w latach trzydziestych ubiegłego stulecia. Pionierem badań w tym zakresie był słynny statystyk angielski Ronald Fisher, który wprowadził pojęcie analizy dyskryminacyjnej (*discriminant analysis*). Zadaniem analizy dyskryminacyjnej jest podział elementów pewnego zbioru na dwie grupy (odpowiadające dwu wartościom rozpatrywanej binarnej zmiennej decyzyjnej), przy czym podział ten jest dokonywany na postawie analizy wartości pozostałych cech statystycznych opisujących rozpatrywane obiekty. Najczęściej stosowanym w praktyce modelem analizy dyskryminacyjnej jest wprowadzona przez Fishera liniowa analiza dyskryminacyjna, w przypadku której o podjętej decyzji klasyfikacyjnej decyduje wartość odpowiednio zdefiniowanej kombinacji liniowej obserwowanych wartości pozostałych cech statystycznych opisujących rozpatrywane obiekty. Warto zauważyć, że znane są także metody nieliniowej (np. kwadratowej) analizy dyskryminacyjnej.

Jak łatwo zauważyć opisany powyżej przykład zadania klasyfikacji jest ściśle związany ze znanym w statystyce od ponad stu lat zagadnieniem wyznaczania zależności regresyjnej. W ujęciu analizy regresyjnej stanowiąca podstawę przyjętej klasyfikacji zmienna binarna (zmienna klasyfikująca) jest zmienną objaśnianą, a pozostałe cechy opisujące klasyfikowane obiekty są zmiennymi objaśniającymi. W takim przypadku wyznaczona na podstawie analizy danych statystycznych funkcja regresji (najczęściej liniowa) służy przewidywaniu wartości zmiennej klasyfikującej.

Historycznie rzecz ujmując funkcje regresji zostały wprowadzone w celu wyznaczenia zależności wartości oczekiwanej zmiennej objaśnianej od wartości zmiennych objaśniających. Jak łatwo zauważyć, wartość oczekiwana zmiennej dyskretnej zazwyczaj nie jest liczbą całkowitą. W związku z tym dla celów klasyfikacji należy wprowadzić dodatkowy warunek pozwalający ustalić wartość dyskretnej zmiennej klasyfikującej na podstawie wyznaczonej oceny jej wartości oczekiwanej. W przypadku zero-jedynkowej klasyfikacji binarnej zazwyczaj przyjmuje się, że wyznaczona ocena wartości oczekiwanej zmiennej klasyfikującej mniejsza od 0,5 skutkuje przyjęciem zerowej wartości zmiennej klasyfikacyjnej, W pozostałym zaś przypadku przyjmujemy, że zmienna klasyfikacyjna ma wartość jeden. Należy jednak pamiętać, że możliwe jest przyjęcie innych rozwiązań, np. gdy koszty błędnych klasyfikacji nie są symetryczne (patrz np. [1]).

W ogólnym przypadku w regresyjnych problemach klasyfikacyjnych można wykorzystywać dowolne funkcje regresji, zarówno liniowe jak i nieliniowe. Na przykład, omawiana w niniejszej pracy *regresja logistyczna* nie jest *sensu stricto* regresją liniową, ale jest przykładem tak zwanego uogólnionego modelu liniowego. W niniejszej pracy ograniczymy się do analizy wyłącznie modeli regresji liniowej, w jej wersji podstawowej oraz w postaci tzw. kwadratowej regresji powierzchni odpowiedzi, a także regresji logistycznej, w przypadku której funkcja logitowa (funkcja wiążąca) ma postać funkcji liniowej lub postać kwadratowej funkcji odpowiedzi. Wybór właśnie takich modeli predykcyjnych warunkowany jest ich popularnością, a także – w przypadku modeli wykorzystujących regresję liniową - powszechną dostępnością, np. w postaci narzędzi analizy danych dostępnych w arkuszach kalkulacyjnych.

Zagadnienie klasyfikacji w swym ogólnym sformułowaniu znacznie wykracza poza obszar związany z analizą regresji oraz analizą dyskryminacyjną. Na przykład, wśród zmiennych objaśniających mogą występować zmienne jakościowe, w przypadku których przypisywanie im wartości liczbowych może budzić uzasadnione kontrowersje. Zagadnieniem budowy klasyfikatorów mogących znaleźć zastosowanie w takich bardzo skomplikowanych przypadkach zajmuje się dział uczenia maszynowego, zwany uczeniem maszynowym pod nadzorem (*supervised machine learning*). W przypadku uczenia maszynowego nie zakładamy żadnego modelu matematycznego opisującego związek zmiennej klasyfikującej ze zmiennymi objaśniającymi (zwanymi w terminologii uczenia maszynowego *atrybutami*). Na podstawie pełnych informacji, obejmujących zarówno wartości zmiennej klasyfikacyjnej jak i wartości zmiennych objaśniających, dotyczących pewnego zbioru obiektów, zwanego zbiorem uczącym (treningowym), budowany jest odpowiedni klasyfikator. Można zauważyć, że wspomniane wcześniej klasyfikatory bazujące na funkcji regresji lub metodach dyskryminacyjnych są szczególnymi przypadkami klasyfikatorów zbudowanych w warunkach uczenia maszynowego pod nadzorem. Oprócz nich dużą popularnością cieszą się inne rodzaje klasyfikatorów, takie jak np. drzewa decyzyjne, sieci bayesowskie lub klasyfikatory jądrowe, a także klasyfikatory wykorzystujące sztuczne sieci neuronowe. Wykorzystanie tego typu klasyfikatorów związane jest z koniecznością posiadania specjalistycznego oprogramowania. Z tego też powodu pominiemy je w przeprowadzonej w niniejszej pracy analizie, pomimo tego, że w wielu z rozpatrywanych w niniejszej pracy przypadków są one bardziej efektywne od klasyfikatorów typu regresyjnego.

Analizie regresji poświęcono tysiące książek i artykułów. Licząca sobie blisko 50 lat i przetłumaczona na język polski fundamentalna monografia Drapera i Smitha [2] liczy blisko 500 stron, a jej ostatnie wydanie (w 1998 r.) w języku angielskim ma już ponad 700 stron. Można więc powiedzieć, że jeśli chodzi o jej podstawowe zastosowania, a takie są przedmiotem niniejszej pracy, wiemy praktycznie wszystko. Pozostają jednak pewne szczegóły, które w konkretnych zastosowaniach mogą mieć istotne znaczenie, a rzadko są przedmiotem badań podstawowych. Najlepszym przykładem jest problem normalności rozkładu prawdopodobieństwa zmiennej objaśnianej oraz ewentualnej zależności zmiennych objaśniających. Zagadnienie to zostało w wyczerpujący sposób przeanalizowane przez statystyków. Inaczej jednak wygląda ta sprawa z punktu widzenia użytkownika, który do swojej dyspozycji ma narzędzia dostępne w popularnym programie typu arkusza kalkulacyjnego, takim jak np. Microsoft Excel. Wyznaczane przez takie narzędzie charakterystyki statystyczne wyznaczonej przez program prostej regresji obliczane są przy założeniu normalności rozkładu zmiennej objaśnianej i wzajemnej niezależności zmiennych objaśniających. W przypadku odstępstwa od tego założenia wszelkie wnioski, dotyczące np. statystycznej istotności wybranych zmiennych objaśniających, mogą być błędne.

Innym problemem jest typ zależności pomiędzy zmienną objaśnianą a zmiennymi objaśniającymi. W przypadku prostej regresji liniowej zakłada się, że jest to zależność liniowa. Jeżeli zależność ta nie jest liniowa, to korzystanie z wyznaczonej przez program prostej linii regresji (można ją zawsze wyznaczyć!) może prowadzić do całkowicie błędnych wniosków. Podobny, choć niedający opisać się w tak prosty sposób, jest rozpatrywany w niniejszej pracy przypadek regresji kwadratowej powierzchni odpowiedzi.

Kolejny problem dotyczy wykorzystania modeli regresyjnych do zagadnień klasyfikacji. Jak już wspomniano, w przypadku zadań binarnej klasyfikacji istnieją metody znacznie lepsze od modeli wykorzystujących proste modele regresyjne. Istotne informacje na ten temat można znaleźć w książce Koronackiego i Ćwika [7] lub fundamentalnej monografii Hastiego i in. [3]. Przegląd interesujących wyników można też znaleźć w dostępnym w Internecie raporcie [1]. Jeżeli jednak zdecydujemy się na stosowanie tych prostych modeli (chociażby ze względu na brak dostępu do odpowiedniego oprogramowania), to ciekawym staje się problem odporności przyjętej metody klasyfikacji na odstępstwa od podstawowych założeń analizy regresji.

Istnieje jednak znacznie poważniejszy problem praktyczny związany z zagadnieniem klasyfikacji. W praktycznie wszystkich pracach na ten temat zakłada się, że zbudowany na podstawie pewnego zbioru danych treningowych klasyfikator będzie stosowany do analizy danych generowanych przez *taki sam mechanizm* (choć niekoniecznie znany) jaki miał zastosowanie w przypadku generacji danych treningowych. W praktyce takie założenie może nie być słuszne. Hryniewicz [4] pokazał, że w przypadku zastosowania metod klasyfikacji do bieżącej kontroli jakości procesów produkcyjnych zmiana parametrów kontrolowanego procesu (np. jego rozregulowanie) może w istotny sposób zmienić jakość użytej metody klasyfikacji. Co gorsza, taka zmiana jakości może zachodzić w kierunku trudnym do przewidzenia. Na przykład, obiektywne pogorszenie się jakości kontrolowanego procesu może, w wyniku zmiany własności stosowanego klasyfikatora, skutkować błędnym wskazaniem poprawy jakości takiego procesu. Z podobnymi problemami możemy spotkać się w przypadku innych zastosowań algorytmów klasyfikacji. Wyobraźmy sobie, na przykład, że pewien algorytm klasyfikacji stosowany jest do oceny podatności pacjentów na zastosowanie pewnej terapii. Jeżeli jednak zostanie on zastosowany w przypadku pacjentów różniących się od tych, którzy stanowili zbiór uczący w procesie konstrukcji klasyfikatora (np. w wyniku zastosowania wobec nich innej, dodatkowej, terapii), to klasyfikacja uzyskana z wykorzystaniem danego klasyfikatora może być błędna. Powyższe przykłady świadczą, że rozpatrywany w niniejszej pracy problem odporności klasyfikatora na odstępstwa od założeń może mieć duże znaczenie praktyczne.

W niniejszej pracy do badania własności klasyfikatorów binarnych wykorzystujących proste modele regresyjne wykorzystano model symulacyjny zaproponowany w pracach Hryniewicza [4], [5]. W pracach tych badano własności szerszej grupy klasyfikatorów (m.in. klasyfikatory wykorzystujące metodę liniowej dyskryminacji oraz klasyfikator C4.5 Quinlana mający charakter drzewa decyzyjnego), ale ograniczono się do analizy niewielkiej (np. tylko 10 zbiorów uczących) grupy przykładów. W niniejszej pracy ograniczymy się wyłącznie do analizy klasyfikatorów wykorzystujących modele regresyjne należące do klasy uogólnionych modeli liniowych, ale będziemy rozpatrywać znacznie szersze spektrum zbiorów uczących i zbiorów testowych. Struktura niniejszej pracy jest następująca. W jej drugim punkcie przedstawimy matematyczny model służący do generacji danych w eksperymentach symulacyjnych. Z kolei, w punkcie trzecim przedstawimy wyniki eksperymentów, których celem było porównanie jakości różnych klasyfikatorów regresyjnych, a także ich odporności na występowanie odstępstw od przyjętego modelu matematycznego. Prace zakończy krótkie jej podsumowanie.

2. Generacja danych w eksperymentach symulacyjnych

2.1. Model matematyczny

W przypadku klasyfikacji binarnej matematycznym modelem opisującym dane jest wielowymiarowy rozkład prawdopodobieństwa wielowymiarowej zmiennej losowej $(X_1,\ldots,X_k Z)$, gdzie zmienne losowe $X_1,\ldots,X_k$ są zmiennymi objaśniającymi o dowolnych rozkładach brzegowych, zaś zmienna klasyfikacyjna Z ma rozkład zero-jedynkowy (Bernouillego). Identyfikacja takiego rozkładu prawdopodobieństwa jest w ogólnym przypadku bardzo trudna, a poza nielicznymi przypadkami szczególnymi w praktyce niemożliwa. Dla potrzeb symulacji danych korzystniejsze jest przyjęcie założenia, że opisane są one wielowymiarową zmienną losową $(X_1,\ldots,X_k T)$ o ciągłych rozkładach brzegowych, przy czym zmienna losowa T powiązana jest ze zmienną klasyfikacyjną Z jakimś warunkiem, na przykład

$$Z = \begin{cases} 1 & , \quad T < t_c \\ 0 & , \quad T \geq t_c' \end{cases} \tag{1}$$

gdzie t_c jest pewną wartością krytyczną. Powyższy model został zaproponowany w pracach Hryniewicza [4],[5] dla przypadku kontroli jakości procesu produkcyjnego, w którym na podstawie obserwacji zmiennych objaśniających $X_1,\ldots,X_k$ dokonuje się predykcji czasu życia T produkowanych obiektów, a następnie klasyfikuje się je na potencjalnie zawodne (gdy $T < t_c$) i potencjalnie niezawodne, w przeciwnym przypadku.

Dalsze uproszczenie modelu można uzyskać przyjmując założenie, że kolejne pary zmiennych objaśniających $(X_i, X_{i+1}), i = 1,\ldots,,k-1$ opisane są dwuwymiarowymi rozkładami prawdopodobieństwa zdefiniowanymi przy pomocy dwuwymiarowych kopuł $C_i\big(F_i(X_i), F_{i+1}(X_{i+1})\big), i = 1,\ldots,k-1$, gdzie $F_1(x_1),\ldots,F_k(x_k)$ są dystrybuantami rozkładów brzegowych zmiennych objaśniających $X_1,\ldots,X_k$. Siłę wzajemnej zależności pomiędzy zmiennymi objaśniającymi będziemy określać za pomocą współczynnika asocjacji τ Kendalla, którego wersję populacyjną dla dowolnej kopuły $C(x,y)$ wyznacza się z zależności [8]

$$\tau(X,Y) = 4\iint_{[0,1]^2} C(x,y)dC(x,y) - 1. \tag{2}$$

Zamiast współczynnika asocjacji τ Kendalla do określenia siły zależności możemy też skorzystać ze współczynnika korelacji rangowej ρ Spearmana. Wyznaczanie wartości tego współczynnika jest jednak dość trudne i w związku z tym w rozpatrywanym w niniejszej pracy modelu będziemy korzystać ze znacznie prostszego do wyznaczenia współczynnika τ Kendalla. Przy okazji warto zaznaczyć, że popularna miara zależności, jaką jest współczynnik korelacji liniowej r Pearsona w ogólnym przypadku nie jest dobrym miernikiem siły zależności. Jest on dobrym miernikiem zależności wyłącznie w przypadku nielicznych rozkładów prawdopodobieństwa, a w szczególności w przypadku wielowymiarowego rozkładu normalnego (gaussowskiego). W przeciwieństwie do miar nieparametrycznych, takich jak współczynniki τ Kendalla lub ρ Spearmana, jego wartość

zależy od postaci rozkładów brzegowych. Co więcej, dla niektórych z popularnych rozkładów prawdopodobieństwa współczynnik korelacji liniowej r Pearsona nie przyjmuje wartości z przedziału [-1,1], co może prowadzić do kłopotów z interpretacją wyników analiz statystycznych (patrz praca Hryniewicza i Karpińskiego [6]).

W przyjętym w niniejszej pracy modelu matematycznym zastosowano kolejne uproszczenie ułatwiające zbudowanie modelu symulacyjnego. Wprowadzone zostały fikcyjne ukryte zmienne losowe $HX_1,\ldots,HX_k$ powiązane z odpowiadającymi im zmiennymi objaśniającymi i odznaczające się tym, że ich wartości mierzone są na tej samej skali co zmienna T. Zależności pomiędzy zmiennymi objaśniającymi i odpowiadającymi im zmiennymi ukrytymi opisane są kopułami $C_{Hi}\left(F_i(X_i), F_{Hi}(HX_i)\right), i = 1,\ldots,k$, przy czym siła zależności pomiędzy zmienną objaśniająca a odpowiadającą jej zmienną ukrytą opisana jest, podobnie jak w omawianym powyżej przypadku, współczynnikiem τ Kendalla. Dodatkowo, przyjęto założenie, że wartość oczekiwana każdej zmiennej ukrytej powiązana jest prostą zależnością liniową z wartością oczekiwaną odpowiadającej jej zmiennej objaśniającej. Ostatnie z przyjętych założeń dotyczy modelu zależności pomiędzy zmiennymi ukrytymi $HX_1,\ldots,HX_k$, a zmienną T, od której wartości zależy wartość zmiennej klasyfikującej Z. W rozpatrywanym w niniejszej pracy modelu przyjęto, podobnie jak w pracach Hryniewicza [4], [5], że jest to silnie nieliniowa (a nawet nieróżniczkowalna) zależność deterministyczna typu *max-min*.

2.2. Opis eksperymentu symulacyjnego

Opisany w poprzednim podpunkcie model matematyczny został zaimplementowany w postaci programu komputerowego służącego do symulacji złożonych zależności pomiędzy zmiennymi objaśniającymi i zmienną klasyfikacyjną. W stworzonej implementacji ograniczono się do przypadku k=4 zmiennych objaśniających $X_1,\ldots,X_4$ i odpowiadających im zmiennych ukrytych $HX_1,\ldots,HX_4$. Jako modele (do wyboru użytkownika) zmiennych objaśniających przyjęto następujące rozkłady prawdopodobieństwa: równomierny, normalny, wykładniczy, Weibulla i logarytmiczno-normalny. Z kolei, jako modele zmiennych ukrytych przyjęto rozkłady: wykładniczy, Weibulla oraz logarytmiczno-normalny. Wybór tych właśnie rozkładów brzegowych motywowany był rozpatrywanym w pracach Hryniewicza [4] [5] zastosowaniem tego modelu, jakim jest predykcja niezawodności.

W komputerowej implementacji opisanego w poprzednim punkcie modelu matematycznego przyjęto, że wszelkie zależności pomiędzy zmiennymi losowymi (objaśniającymi i ukrytymi) będą opisywane następującymi kopułami (w terminologii polskiej stosowany jest też termin kopula):

- niezależności

$$C(u,v) = u * v;\ u,v \in [0,1], \tag{3}$$

- Claytona

$$C(u,v) = \left[max\left(u^{-\theta} + v^{-\theta} - 1; 0\right)\right]^{-1/\theta}, \theta \in [-1,\infty)\backslash\{0\}; u,v \in [0,1], \tag{4}$$

- Gumbela

$$C(u,v) = exp\left[-\left(log(u)\right)^{\theta} + \left(-log(v)\right)^{\theta}\right)^{1/\theta}\right], \theta \in [1,\infty); u,v \in [0,1], \quad (5)$$

- Franka

$$C(u,v) = -\frac{1}{\theta} log\left[1 + \frac{(exp(-\theta u)-1)(exp(-\theta v)-1)}{(exp(-\theta)-1)}\right], \theta \in \boldsymbol{R}\backslash\{0\};\ u,v \in [0,1], \quad (6)$$

- Fairlie-Gumbela-Morgensterna (FGM)

$$C(u,v) = uv[1 + \theta(1-u)(1-v)], \theta \in [-1,1];\ u,v \in [0,1] \quad (7)$$

- gaussowską (normalną)

$$C(u,v) = \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); r), r \in [-1,1];\ u,v \in [0,1], \quad (8)$$

gdzie $\Phi_2(x,y;r)$ jest dystrybuantą dwuwymiarowego rozkładu normalnego (Gaussa) o współczynniku korelacji r, zaś $\Phi^{-1}(x)$ jest funkcją odwrotną (kwantylową) dystrybuanty jednowymiarowego rozkładu normalnego.

Siła zależności pomiędzy zmiennymi losowymi powiązanymi powyższymi kopułami określona została przy pomocy współczynnika τ Kendalla, zdefiniowanego wzorem (2). Odpowiednie wzory na wartość tego współczynnika można znaleźć w fundamentalnej monografii Nelsena [8], a także w pracy Hryniewicza i Karpińskiego [6].

Opisany powyżej model symulacyjny może służyć do badania własności różnych zadań klasyfikacji. W opisanym w następnym punkcie niniejszej pracy eksperymencie analizie poddano sześć regresyjnych modeli klasyfikacji:

- binarna regresja liniowa 4 zmiennych o funkcji regresji danej wzorem

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4, \quad (9)$$

- regresja liniowa 4 zmiennych zmiennej pomocniczej T

$$t = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_4, \quad (10)$$

- regresja kwadratowej powierzchni odpowiedzi zmiennej binarnej

$$y_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 {x_1}^2 + \beta_6 {x_2}^2 + \beta_7 {x_3}^2 + \beta_8 {x_4}^2 + \beta_9 x_1 x_2 + \beta_{10} x_1 x_3 + \beta_{11} x_1 x_4 + \beta_{12} x_2 x_3 + \beta_{13} x_2 x_4 + \beta_{14} x_3 x_4, \quad (11)$$

- regresja kwadratowej powierzchni odpowiedzi zmiennej pomocniczej T

$$t_2 = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_4 + \gamma_5 {x_1}^2 + \gamma_6 {x_2}^2 + \gamma_7 {x_3}^2 + \gamma_8 {x_4}^2 + \gamma_9 x_1 x_2 + \gamma_{10} x_1 x_3 + \gamma_{11} x_1 x_4 + \gamma_{12} x_2 x_3 + \gamma_{13} x_2 x_4 + \gamma_{14} x_3 x_4. \quad (12)$$

- regresja logistyczna z liniową funkcją wiążącą

$$ln\left(\frac{p}{1-p}\right) = \delta_0 + \delta_1 x_1 + \delta x_2 + \delta_3 x_3 + \delta_4 x_4 \quad (13)$$

- regresja logistyczna z funkcją wiążącą w postaci kwadratowej funkcji odpowiedzi

$$ln\left(\frac{p}{1-p}\right) = \sigma_0 + \sigma_1 x_1 + \sigma_2 x_2 + \sigma_3 x_3 + \sigma_4 x_4 + \sigma_5 {x_1}^2 + \sigma_6 {x_2}^2 + \sigma_7 {x_3}^2 + \sigma_8 {x_4}^2 + \sigma_9 x_1 x_2 + \sigma_{10} x_1 x_3 + \sigma_{11} x_1 x_4 + \sigma_{12} x_2 x_3 + \sigma_{13} x_2 x_4 + \sigma_{14} x_3 x_4. \quad (14)$$

Parametry modeli liniowych (9)–(12) estymowane są metodą najmniejszych kwadratów, a parametry modeli logistycznych (12)–(13) estymowane są metodą największej wiarogodności.

W eksperymentach symulacyjnych najpierw generowano próbki uczące (treningowe), na podstawie których identyfikowano powyższe modele regresyjne. Modele (9) i (11) budowane były przy założeniu, że zmienna objaśniana ma postać binarną, zaś modele (10) i (12) budowane były dla pomocniczej zmiennej T o wartościach rzeczywistych.

Następnie, generowane były próbki testowe, w przypadku których wyniki predykcji, uzyskane za pomocą równań regresji (9) – (12) były przekształcane na wartości zmiennej klasyfikującej (0 lub 1). Wyniki klasyfikacji były oceniane na podstawie analizy następującej macierzy:

Tabela 1. Podsumowanie wyników klasyfikacji binarnej

		Rzeczywista klasa	
		1	0
Przewidywana klasa	1	TP	FP
	0	FN	TN

Symbole opisujące komórki Tabeli 1 odpowiadają powszechnie przyjętej w uczeniu maszynowym terminologii (przejętej z analizy danych medycznych). Symbole TP (*True Positives*) oraz TN (*True Negatives*) oznaczają liczby poprawnie zaklasyfikowanych obiektów (odpowiednio, jedynek oraz zer). Symbol FP (*False Positives*) oznacza liczbę zer błędnie sklasyfikowanych jako jedynki, zaś symbol FN oznacza liczbę jedynek błędnie sklasyfikowanych jako zera. Na podstawie tak określonych danych określa się różne miary jakości klasyfikacji.

Podstawowa miara jakości klasyfikacji nazywana jest *Dokładność* (*Accuracy*) i jest definiowana w następujący sposób:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \tag{15}$$

Jest to więc frakcja jednostek poprawnie sklasyfikowanych. Miara ta byłaby wystarczająca gdyby nie miało znaczenia, które jednostki (jedynki lub zera) klasyfikujemy poprawnie, a które niepoprawnie. Żeby uwzględnić możliwą (i bardzo często występującą w praktyce) asymetrię konsekwencji wyników klasyfikacji wprowadzono dodatkowe wskaźniki jakości, takie jak *Precyzja* (*Precision*)

$$Precision = \frac{TP}{TP+FP}, \tag{16}$$

Czułość (*Sensitivity*)

$$Sensitivity = \frac{TP}{TP+FN}, \tag{17}$$

oraz *Specyficzność* (*Specificity*)

$$Specificity = \frac{TN}{TN+FP}. \tag{18}$$

Stosowane są także pewne wskaźniki kompleksowe, takie jak np. *Indeks F1* (*F1-Index*)

$$F1 = \frac{2TP}{2TP+FP+FN}. \tag{19}$$

Szczegółową interpretację niektórych z powyższych wskaźników, a także opis innych wskaźników jakości klasyfikacji można znaleźć m.in. w książce Koronackiego i Ćwika [7] oraz monografii [3] .

Ogólnie rzecz biorąc, klasyfikatory odznaczające się wyższą jakością powinny mieć wyższe wartości powyższych wskaźników. Odpowiednia wersja znanego „twierdzenia o darmowych obiadach" Wolperta mówi nam jednak, że nie istnieje klasyfikator, który dawałby lepsze rezultaty klasyfikacji dla wszystkich możliwych danych. Należy się więc spodziewać, że żaden z rozpatrywanych w niniejszej pracy liniowych klasyfikatorów regresyjnych nie będzie wykazywał przewagi nad pozostałymi dla wszystkich stosowanych do porównań miar jakości klasyfikacji. Przykłady numeryczne, opisane w następnym punkcie niniejszej pracy, potwierdzają to przypuszczenie.

3. Wyniki eksperymentów symulacyjnych

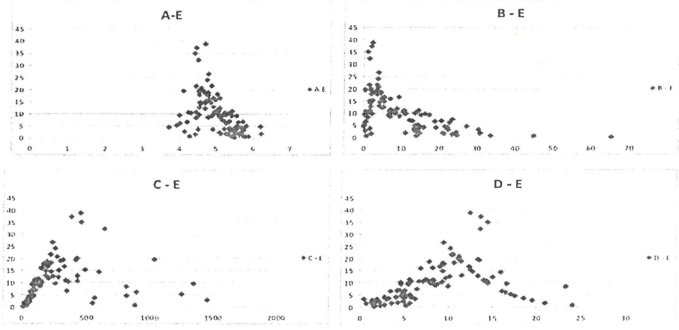
Eksperymenty symulacyjne mające na celu porównanie jakości rozpatrywanych w niniejszej pracy klasyfikatorów regresyjnych prowadzono dla różnych kombinacji takich wielkości jak: rozkłady prawdopodobieństwa zmiennych objaśniających $X_1,\ldots,X_4$ i odpowiadających im zmiennych ukrytych $HX_1,\ldots,HX_4$, kopuły opisujące wzajemne zależności pomiędzy zmiennymi objaśniającymi oraz pomiędzy zmiennymi objaśniającymi oraz odpowiadającymi im zmiennymi ukrytymi, a także wartości miar siły zależności pomiędzy rozpatrywanymi zmiennymi. Uwzględniono również różne zależności regresyjne pomiędzy wartościami oczekiwanymi zmiennych objaśniających i odpowiadających im zmiennych ukrytych, a także różne postacie funkcji wiążących wartości zmiennych ukrytych i ciągłej zmiennej wyjściowej T. W niniejszej pracy zaprezentowano jedynie wyniki uzyskane dla modelu identycznego do tego, jaki był opisany w pracach [4] oraz [5]. Pozwoli to odnosić się to rezultatów opisanych w tych pracach i w rezultacie ograniczyć objętość niniejszego artykułu.

W rozpatrywanym konkretnym modelu symulacyjnym służącym do konstrukcji klasyfikatorów przyjęto, że zmienna objaśniająca X_1 ma rozkład normalny o wartości oczekiwanej 5 i odchyleniu standardowym 0,5, zmienna X_2 ma rozkład wykładniczy o wartości oczekiwanej 10, zmienna X_3 ma rozkład logarytmiczno-normalny o wartości oczekiwanej jej logarytmu 5 i odchyleniu standardowym jej logarytmu równym 1, zaś zmienna X_4 ma rozkład Weibulla o parametrze skali 10 i parametrze kształtu 2. Dla zmiennych ukrytych przyjęto następujące rozkłady: HX_1 – logarytmiczno-normalny, HX_2 – wykładniczy, HX_3 – wykładniczy, HX_2 – Weibulla o parametrze kształtu 1,5.

Do opisu zależności pomiędzy parami zmiennych użyto następujących kopuł (w nawiasie wartość współczynnika τ Kendalla: (X_1,X_2) – Clayton (0,8), (X_2,X_3) – gaussowska (-0,8), (X_3,X_4) – Frank (0,8), (X_1,HX_1) – gaussowska (-0,8), (X_2,HX_2) – Frank (0,9), (X_3,HX_3) – Gumbel (0,9), (X_4,HX_4) – Clayton (-0,8). Związek pomiędzy zmienną wyjściową T a zmiennymi ukrytymi określono zależnością nieliniową T = min [max $(HX_1; HX_2)$; min$(HX_3;HX_4)$]. Warto zauważyć, że np. w przypadku zmiennych objaśniających (X_2,X_3) ich

dwuwymiarowy rozkład prawdopodobieństwa jest rozkładem „normalnym" o rozkładach brzegowych niebędących rozkładami normalnymi (i na dodatek różnymi). Taką elastyczność modelowania złożonych zjawisk stochastycznych można uzyskać wykorzystując formalizm kopuł.

W pracy [4] podano wykres zależności pomiędzy zmiennymi objaśniającymi (oznaczonymi A, B, C D), a zmienną wyjściową (oznaczoną E) dla przykładowego zbioru danych wygenerowanych za pomocą opisanego powyżej modelu.



Rysunek 1. Przykładowe zależności pomiędzy wartościami zmiennych objaśniających i zmiennej objaśnianej [4]

Jak widać, zależności te są silnie nieliniowe oraz niemonotoniczne. W tej sytuacji powstaje zasadnicze pytanie, czy proste zależności regresyjne są przydatne w zagadnieniach klasyfikacji danych opisanych tak nietypowymi złożonymi modelami. W tym celu przeprowadzono eksperyment symulacyjny, w którym wygenerowano 50 zbiorów danych treningowych o licznościach, odpowiednio, 100, 200, 300 oraz 500 obserwacji. Dla każdego z tych zbiorów wyznaczono sześć modeli regresyjnych:

- liniową regresję binarną dla 4 zmiennych objaśniających (REG-4),
- liniową regresję dla zmiennej pośredniej T i 4 zmiennych objaśniających (REG-T4),
- binarną regresję kwadratowej powierzchni odpowiedzi i 4 zmiennych objaśniających (REG-14),
- regresję kwadratowej powierzchni odpowiedzi dla zmiennej pośredniej T i 4 zmiennych objaśniających (REG-T14),
- regresję logistyczną z liniową funkcją wiążącą (*link function*) 4 zmiennych objaśniających (LOG-REG4) oraz
- regresję logistyczną z funkcją wiążącą w postaci kwadratowej funkcji odpowiedzi dla 4 zmiennych objaśniających (LOG-REG14).

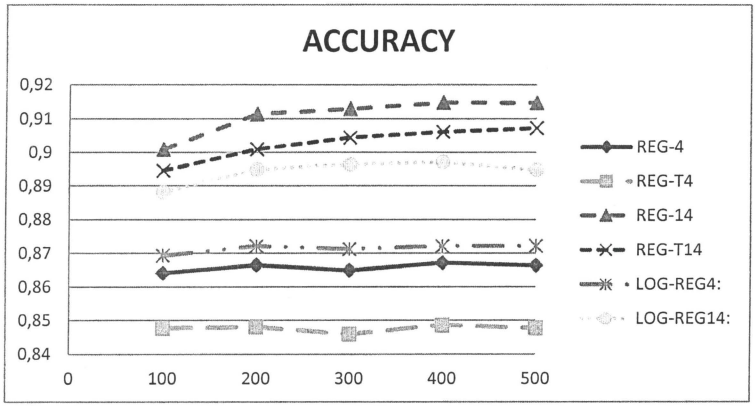
Następnie dla każdego zbioru danych treningowych wygenerowano 50 zbiorów danych testowych o liczności 1000 elementów, na których weryfikowano jakość otrzymanych (dla danego zbioru treningowego) klasyfikatorów.

Wyniki opisanego powyżej eksperymentu przedstawiono w Tabelach 2-6 oraz na Rysunkach 2 – 6. Z ich analizy wynika, że żaden z porównywanych klasyfikatorów w sposób wyraźny nie ma przewagi nad pozostałymi. Biorąc jednak pod uwagę wartości wyznaczonych wskaźników jakości można przyjąć, że klasyfikatory wykorzystujące regresję kwadratowej

powierzchni, uwzględniające efekt interakcji pomiędzy zmiennymi objaśniającymi, odpowiedzi są nieco lepsze od klasyfikatorów wykorzystujących prostą regresję liniową.

Tabela 2. Porównanie *Dokładności* klasyfikatorów regresyjnych

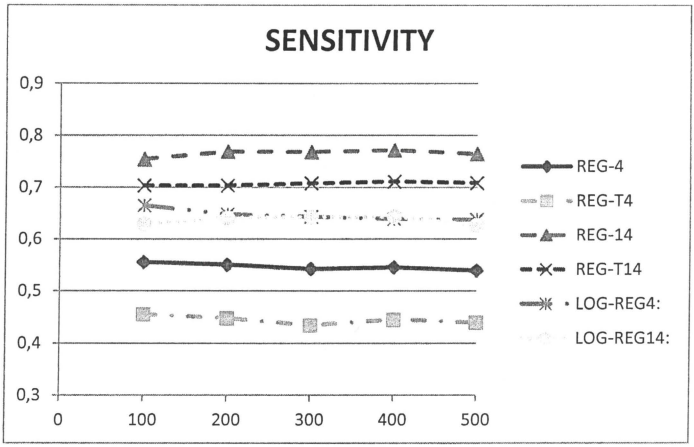
n	REG-4	REG-T4	REG-14	REG-T14	LOG-REG4	LOG-REG14
100	0,8641	0,8477	0,9009	0,8946	0,8693	0,8881
200	0,8665	0,8481	0,9113	0,9010	0,8721	0,8949
300	0,8649	0,8459	0,9128	0,9044	0,8712	0,8964
400	0,8671	0,8486	0,9147	0,9061	0,8721	0,8971
500	0,8664	0,8477	0,9146	0,9072	0,8722	0,8947



Rysunek 2. Porównanie *Dokładności* klasyfikatorów regresyjnych

Tabela 3. Porównanie *Czułości* klasyfikatorów regresyjnych

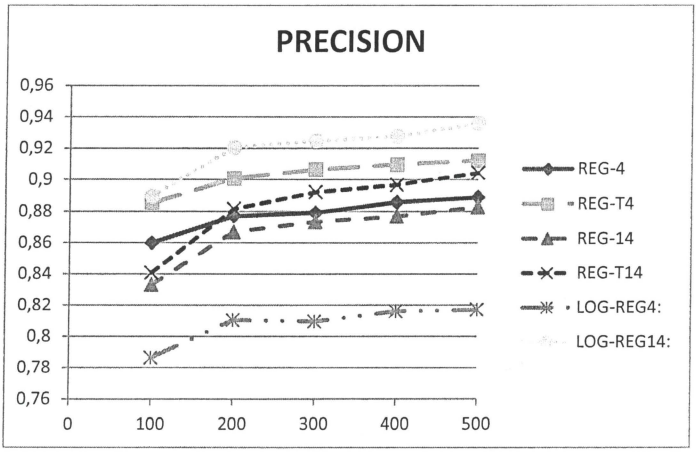
n	REG-4	REG-T4	REG-14	REG-T14	LOG-REG4	LOG-REG14
100	0,5556	0,4555	0,7539	0,7034	0,6648	0,6295
200	0,5505	0,4482	0,7685	0,7032	0,6480	0,6409
300	0,5422	0,4345	0,7678	0,7075	0,6434	0,6429
400	0,5453	0,4454	0,7714	0,7108	0,6390	0,6436
500	0,5393	0,4395	0,7643	0,7081	0,6376	0,6268



Rysunek 3. Porównanie *Czułości* klasyfikatorów regresyjnych

Tabela 4. Porównanie *Precyzji* klasyfikatorów regresyjnych

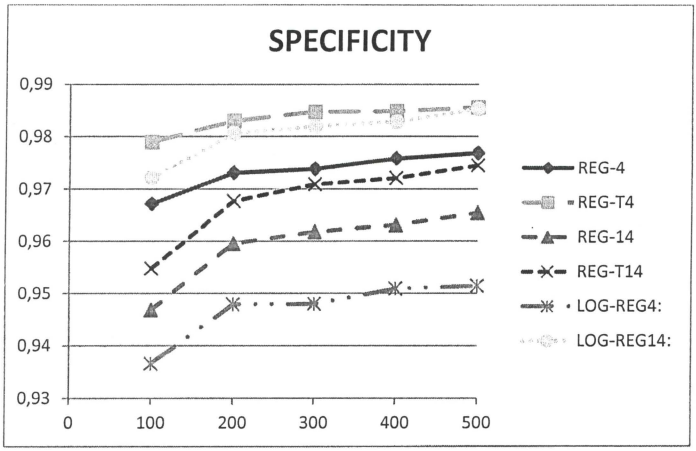
n	REG-4	REG-T4	REG-14	REG-T14	LOG-REG4	LOG-REG14
100	0,8599	0,8853	0,8330	0,8406	0,7863	0,8896
200	0,8769	0,9007	0,8669	0,8815	0,8104	0,9203
300	0,8790	0,9064	0,8735	0,8920	0,8095	0,9246
400	0,8856	0,9095	0,8768	0,8966	0,8161	0,9275
500	0,8887	0,9120	0,8827	0,9042	0,8170	0,9358



Rysunek 4. Porównanie *Precyzji* klasyfikatorów regresyjnych

Tabela 5. Porównanie *Specyficzności* klasyfikatorów regresyjnych

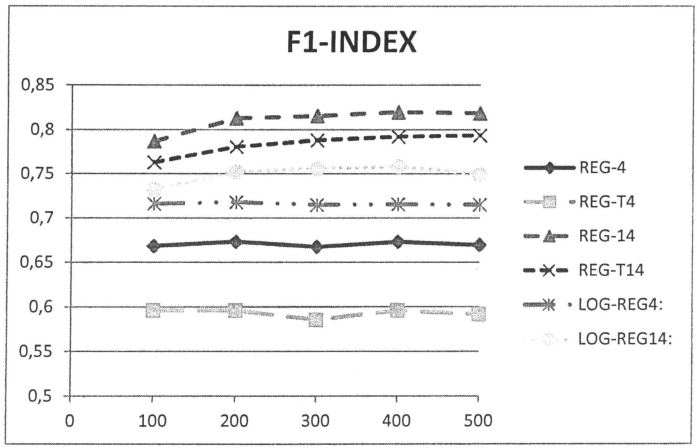
n	REG-4	REG-T4	REG-14	REG-T14	LOG-REG4	LOG-REG14
100	0,9672	0,9790	0,9468	0,9548	0,9366	0,9722
200	0,9731	0,9830	0,9596	0,9677	0,9479	0,9806
300	0,9738	0,9847	0,9618	0,9709	0,948	0,9819
400	0,9758	0,9848	0,9631	0,9721	0,9509	0,9828
500	0,9769	0,9856	0,9654	0,9745	0,9514	0,9853



Rysunek 5. Porównanie *Specyficzności* klasyfikatorów regresyjnych

Tabela 6. Porównanie *indeksu F1* klasyfikatorów regresyjnych

n	REG-4	REG-T4	REG-14	REG-T14	LOG-REG4	LOG-REG14
100	0,6684	0,5963	0,7868	0,7626	0,7159	0,7319
200	0,6732	0,5963	0,8128	0,7806	0,7179	0,7521
300	0,6674	0,5857	0,8154	0,7880	0,7148	0,7558
400	0,6732	0,5962	0,8196	0,7920	0,7154	0,7581
500	0,6696	0,5919	0,8183	0,7934	0,7151	0,7491



Rysunek 6. Porównanie *indeksu F1* klasyfikatorów regresyjnych

Powyższy wniosek potwierdza analiza „rang" poszczególnych klasyfikatorów uzyskanych w procesie ich porównania na przyjętym zbiorze wskaźników jakości. Przyjmujemy, że jeden klasyfikator ma niższą (lepszą!) rangę od drugiego, gdy odpowiadający mu wykres znajduje się wyżej na Rysunkach 2 – 6. Podsumowanie takiego porównania przedstawiono w Tabeli 7.

Tabela 7. Porównanie jakości klasyfikatorów

Wskaźnik jakości	REG-4	REG-T4	REG-14	REG-T14	LOG-REG4	LOG-REG14
Dokładność	5	6	1	2	4	3
Czułość	5	6	1	2	3	4
Precyzja	4	2	5	3	6	1
Specyficzność	3	1	5	4	6	2
F1	5	6	1	2	4	3

Uśrednione wartości rang klasyfikatorów regresyjnych wykorzystujących, jako funkcję wiążącą, kwadratową funkcję odpowiedzi (REG-14 REG-T14 oraz LOG-REG14) są jednakowe i wynoszą 2,6. W przypadku dodatkowego premiowania „pierwszych miejsc" najlepszym staje się klasyfikator REG-14. Klasyfikator ten wypada też najlepiej w przypadku dokonania bezpośrednich porównań tych trzech klasyfikatorów. Z kolei, w przypadku klasyfikatorów z liniową funkcją wiążącą (REG-4, REG-T4 oraz LOG-REG4) średnie rangi są bardzo podobne z niewielką przewagą klasyfikatora REG-T4. Jednakże w przypadku gdy porównamy wyłącznie te trzy klasyfikatory najlepszym wydaje się być klasyfikator LOG-REG4, bazujący na prostej regresji logistycznej. Pierwszym wnioskiem z tego porównania jest to, że klasyfikatory regresyjne wykorzystujące bardziej złożoną funkcję wiążącą są bardziej efektywne od klasyfikatorów wykorzystujących prostą funkcję liniową. Można też wyciągnąć wniosek, że jakość klasyfikatorów bazujących na danych binarnych nie odbiega od jakości klasyfikatorów budowanych na podstawie obserwacji zmiennych pośrednich o wartościach rzeczywistych, co nie wydawało się być rzeczą oczywistą. Z analizy przedstawionych powyżej rysunków wynika też, że dla przyjętego zakresu liczności zbiorów

uczących jakość klasyfikatorów niewiele się zmienia. Jedynie w przypadku klasyfikatorów bazujących na regresji kwadratowej powierzchni odpowiedzi jakość ta rośnie wraz ze wzrostem liczności zbioru uczącego, co nie jest dziwne, biorąc pod uwagę fakt, że estymowany model ma aż 15 parametrów.

Opisany powyżej eksperyment symulacyjny dotyczył przypadku gdy model matematyczny opisujący dane uczące (treningowe), na podstawie których budowany jest klasyfikator i model matematyczny danych, które poddawane są procesowi klasyfikacji za pomocą tego klasyfikatora, są takie same. Jest to podstawowe założenie stosowane w uczeniu maszynowym. Jeżeli jednak myślimy o praktycznym zastosowaniu danego konkretnego klasyfikatora, to musimy określić jego odporność na zmiany modelu opisującego dane. Należy podkreślić, że problem ten w zasadniczy sposób różni się od tego czym jest zainteresowane środowisko naukowe zajmujące się problemami klasyfikacji i, szerzej, uczenia maszynowego. Celem prowadzonych badań z tej dziedziny jest znalezienie takiej metody konstrukcji klasyfikatora, by można ją było z powodzeniem zastosować dla różnych zbiorów danych. Tymczasem dla osoby stosującej metody klasyfikacji w swej pracy ważne jest by konkretny klasyfikator, często wyznaczony z dużym nakładem sił i środków, spełniał swoje zadanie w ulegającym zmianie środowisku.

W pracach Hryniewicza [4] i [5] rozpatrywany był problem klasyfikacji produkowanych wyrobów na potencjalnie zawodne i potencjalnie niezawodne. W przypadku stabilnych procesów produkcyjnych raz skonstruowany klasyfikator może być efektywnie wykorzystywany przez dłuższy czas. Sytuacja ulegnie jednak zmianie, gdy wystąpią zakłócenia kontrolowanego procesu. Z przykładów omawianych w tych pracach wynika, że nawet niewielkie zmiany parametrów procesu, które nie mają praktycznie żadnego wpływu na jego jakość, mogą w radykalny sposób pogorszyć jakość klasyfikacji, co z kolei może prowadzić do powstawania fałszywych alarmów. Sytuacja może być jednak jeszcze gorsza. W pewnych przypadkach zmiany parametrów procesu prowadzące do pogorszenia się jego jakości mogą skutkować takimi zmianami jakości klasyfikacji, że kontrolowany proces może – całkowicie błędnie – wykazywać symptomy poprawy a nie pogorszenia, jak jest w rzeczywistości. Przeanalizowanie wpływu takich niestabilności na jakość omawianych w niniejszej pracy klasyfikatorów było celem następnego eksperymentu symulacyjnego.

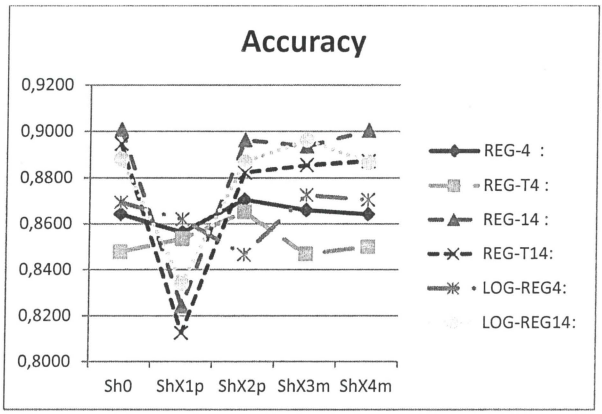
W pracach [4] i [5] pokazano, że dla przyjętego modelu generacji danych zmiany rozkładów prawdopodobieństwa zmiennych objaśniających $X_1, \ldots, X_4$ mogą mieć bardzo różne konsekwencje. Okazuje się, że zwiększenie wartości oczekiwanych zmiennych objaśniających X_1 oraz X_2 prowadzi do zwiększenia się frakcji jednostek zaklasyfikowanych do klasy „1". Z kolei, do wystąpienia tego samego efektu prowadzi zmniejszenie się wartości oczekiwanych zmiennych objaśniających X_3 oraz X_4. Dodatkowo okazało się, że z praktycznego punktu widzenia zmiany frakcji jednostek zaklasyfikowanych do klasy „1", które były wynikiem wzrostu wartości oczekiwanych zmiennych objaśniających X_1 oraz X_2 są pomijalne. Natomiast analogiczne zmiany będące efektem zmniejszenia się wartości oczekiwanych zmiennych objaśniających X_3 oraz X_4 mają istotne znaczenie. W takim przypadku każda zmiana jakości klasyfikatora związana ze wzrostem wartości oczekiwanych zmiennych objaśniających X_1 oraz X_2 będzie prowadzić do zwiększenia częstości fałszywych sygnałów wskazujących na zmianę parametrów kontrolowanego procesu. W przypadku zmian zmiennych X_3 oraz X_4 sytuacja jest bardziej skomplikowana, gdyż zmiana jakości

klasyfikatora może skutkować np. szybszym wykryciem rozregulowania się kontrolowanego procesu. Dokładna ocena skutków takiej zmiany jakości klasyfikatora wymaga w takich przypadkach przeprowadzenia złożonych analiz techniczno-ekonomicznych. Z powyższych rozważań wynika jednak, że pożądaną własnością klasyfikatora jest w każdym z omawianych tu przypadków odpowiednia odporność na zmiany parametrów procesu.

W omówionym poniżej eksperymencie symulacyjnym przyjęto, że zmiana parametrów procesu polega na zwiększeniu się wartości oczekiwanych zmiennych objaśniających X_1 oraz X_2 o wartość równą 0,5 ich odchylenia standardowego oraz analogicznym zmniejszeniu się wartości oczekiwanych zmiennych objaśniających X_3 oraz X_4. Na wykresach przedstawiających wyniki symulacji zmienione wartości parametrów zmiennych objaśniających oznaczono, odpowiednio, przez ShX1p, ShX2p, ShX3m, ShX4m, zaś przypadek procesu stabilnego oznaczono przez Sh0. Przyjęto też dodatkowe założenie, że zmianie ulega wartość oczekiwana tylko jednej ze zmiennych objaśniających. Wyniki eksperymentu przedstawiono w Tabelach 8 – 12 oraz na Rysunkach 7 – 11.

Tabela 8. Porównanie *Dokładności* klasyfikatorów regresyjnych dla zmiennych parametrów modelu

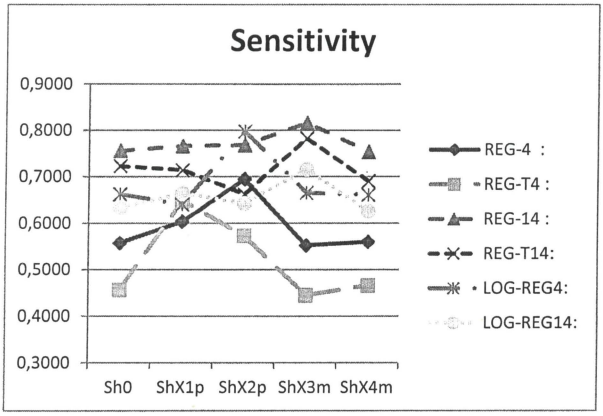
	Sh0	ShX1p	ShX2p	ShX3m	ShX4m
REG-4 :	0,8641	0,8562	0,8704	0,8659	0,8641
REG-T4 :	0,8477	0,8534	0,8651	0,8468	0,8499
REG-14 :	0,9009	0,8241	0,8962	0,8935	0,9003
REG-T14:	0,8946	0,8125	0,8822	0,8854	0,8871
LOG-REG4:	0,8693	0,8619	0,8464	0,8725	0,8704
LOG-REG14:	0,8881	0,8340	0,8868	0,8962	0,8866



Rysunek 7. Porównanie *Dokładności* klasyfikatorów regresyjnych dla zmiennych parametrów modelu

Tabela 9. Porównanie *Czułości* klasyfikatorów regresyjnych dla zmiennych parametrów modelu

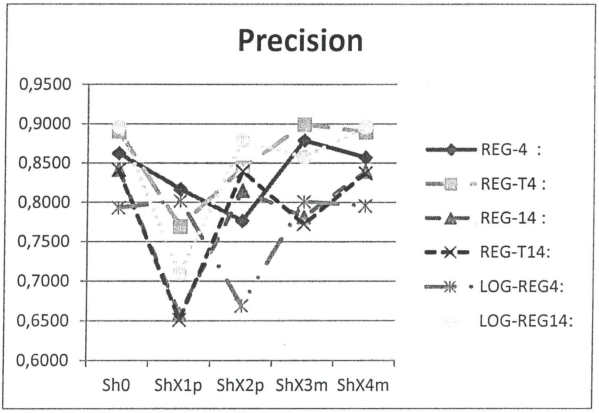
	Sh0	ShX1p	ShX2p	ShX3m	ShX4m
REG-4 :	0,5571	0,6035	0,6943	0,5520	0,5598
REG-T4 :	0,4561	0,6482	0,5732	0,4444	0,4654
REG-14 :	0,7553	0,7656	0,7681	0,8155	0,7532
REG-T14:	0,7222	0,7135	0,6620	0,7820	0,6895
LOG-REG4:	0,6618	0,6405	0,7974	0,6651	0,6607
LOG-REG14:	0,6335	0,6638	0,6426	0,7153	0,6258



Rysunek 8. Porównanie *Czułości* klasyfikatorów regresyjnych dla zmiennych parametrów modelu

Tabela 10. Porównanie *Precyzji* klasyfikatorów regresyjnych dla zmiennych parametrów modelu

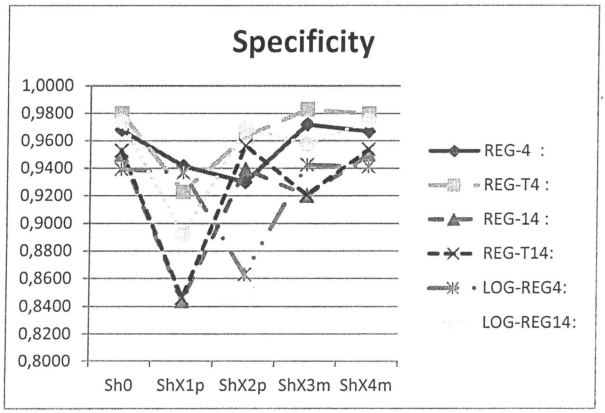
	Sh0	ShX1p	ShX2p	ShX3m	ShX4m
REG-4 :	0,8621	0,8169	0,7765	0,8786	0,8569
REG-T4 :	0,8903	0,7693	0,8441	0,8990	0,8895
REG-14 :	0,8415	0,6583	0,8152	0,7815	0,8395
REG-T14:	0,8411	0,6515	0,8393	0,7723	0,8370
LOG-REG4:	0,7935	0,8023	0,6691	0,8010	0,7957
LOG-REG14:	0,8965	0,7097	0,8789	0,8574	0,8959



Rysunek 9. Porównanie *Precyzji* klasyfikatorów regresyjnych dla zmiennych parametrów modelu

Tabela 11. Porównanie *Specyficzności* klasyfikatorów regresyjnych dla zmiennych parametrów modelu

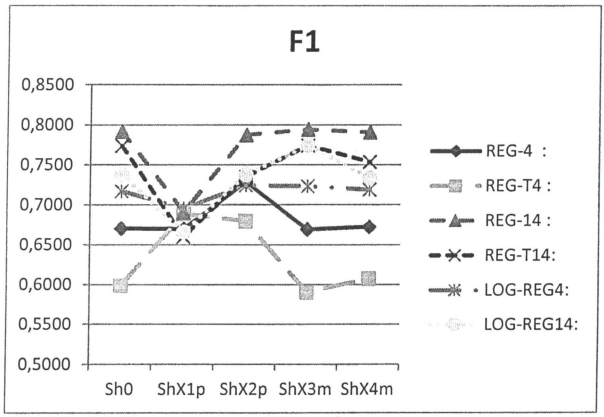
	Sh0	ShX1p	ShX2p	ShX3m	ShX4m
REG-4 :	0,9678	0,9415	0,9297	0,9720	0,9668
REG-T4 :	0,9799	0,9227	0,9634	0,9828	0,9797
REG-14 :	0,9500	0,8439	0,9393	0,9199	0,9500
REG-T14:	0,9528	0,8460	0,9564	0,9204	0,9539
LOG-REG4:	0,9392	0,9366	0,8629	0,9425	0,9412
LOG-REG14:	0,9741	0,8914	0,9690	0,9574	0,9747



Rysunek 10. Porównanie *Specyficzności* klasyfikatorów regresyjnych dla zmiennych parametrów modelu

Tabela 12. Porównanie *indeksu F1* klasyfikatorów regresyjnych dla zmiennych parametrów modelu

	Sh0	ShX1p	ShX2p	ShX3m	ShX4m
REG-4 :	0,6704	0,6700	0,7286	0,6694	0,6724
REG-T4 :	0,5983	0,6889	0,6796	0,5904	0,6077
REG-14 :	0,7918	0,6907	0,7874	0,7942	0,7907
REG-T14:	0,7738	0,6602	0,7352	0,7742	0,7535
LOG-REG4:	0,7170	0,6953	0,7243	0,7231	0,7188
LOG-REG14:	0,7376	0,6660	0,7377	0,7743	0,7335



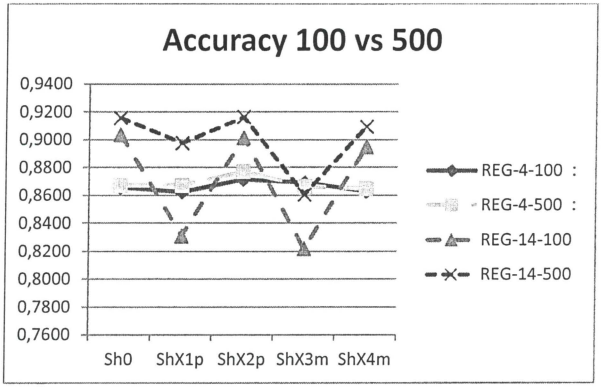
Rysunek 11. Porównanie *indeksu F1* klasyfikatorów regresyjnych dla zmiennych parametrów modelu

Z analizy powyższych wykresów wynika, że klasyfikatory wykorzystujące prosty model regresji binarnej, a także prostej regresji logistycznej, zachowują się bardziej stabilnie. W przypadku klasyfikatorów wykorzystujących model regresji kwadratowej powierzchni odpowiedzi zmiana wartości oczekiwanej zmiennych objaśniających X_1 oraz X_3 prowadzi do radykalnej zmiany jakości klasyfikacji, przy czym tylko w przypadku zmiany *Czułości* jest to zmiana na lepsze. Z porównań wynika również to, że klasyfikatory wykorzystujące zmienne binarne są bardziej odporne na zmiany parametrów modelu niż klasyfikatory wykorzystujące wartości ciągłej zmiennej pośredniej *T*.

W omówionym powyżej eksperymencie przyjęto, że liczność zbioru uczącego wynosi 100 elementów. Interesującym może być pytanie czy i jak zmienią się jakościowe charakterystyki klasyfikatora w przypadku zastosowania większych zbiorów uczących. W Tabeli 13 oraz na Rysunku 12 przedstawiono porównanie *Dokładności* klasyfikatorów Reg-4 oraz Reg-14 dla przypadku zbiorów uczących o liczności, odpowiednio, 100 i 500 elementów.

Tabela 13. Porównanie *Dokładności* liniowych klasyfikatorów regresyjnych dla zmiennych parametrów modelu dla różnych liczności zbiorów uczących

	Sh0	ShX1p	ShX2p	ShX3m	ShX4m
REG-4-100	0,8655	0,8622	0,8713	0,8690	0,8628
REG-4-500	0,867	0,8672	0,8771	0,8665	0,8646
REG-14-100	0,9035	0,8310	0,9014	0,8217	0,8950
REG-14-500	0,9157	0,8975	0,9163	0,8604	0,9093

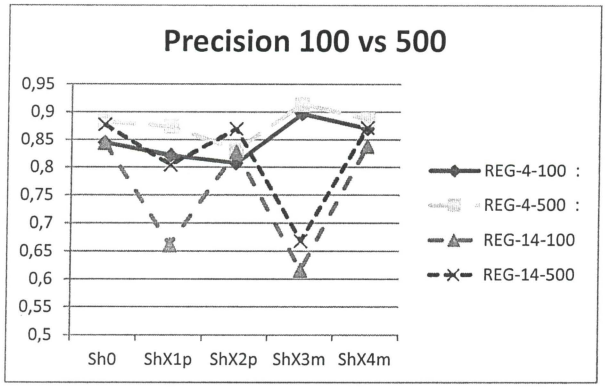


Rysunek 12. Porównanie *Dokładności* liniowych klasyfikatorów regresyjnych dla zmiennych parametrów modelu dla różnych liczności zbiorów uczących

Z Rysunku 12 wynika, że porównywane liczności zbiorów uczących nie mają wpływu na dokładność prostych klasyfikatorów binarnych. Natomiast w przypadku klasyfikatora binarnego wykorzystującego regresję kwadratowej powierzchni odpowiedzi zwiększenie liczności zbioru uczącego nie tylko zwiększa dokładność klasyfikacji, ale także zwiększa stabilność klasyfikatora. Podobny efekt uzyskujemy gdy porównamy precyzję klasyfikatorów uzyskanych dla zbiorów uczących o różnych licznościach. W tym przypadku, przedstawionym w Tabeli 14 i na Rysunku 13, zwiększenie liczności zbioru uczącego ze 100 do 500 poprawiło precyzję klasyfikatora binarnego wykorzystującego regresję kwadratowej powierzchni odpowiedzi oraz zwiększyło jego stabilność. Nie zmienia to faktu, że klasyfikatory wykorzystujące prostą regresję binarną są nadal, biorąc pod uwagę tę miarę jakości klasyfikacji, bardziej stabilne.

Tabela 14. Porównanie *Precyzji* liniowych klasyfikatorów regresyjnych dla zmiennych parametrów modelu dla różnych liczności zbiorów uczących

	Sh0	ShX1p	ShX2p	ShX3m	ShX4m
REG-4-100	0,8446	0,8218	0,8075	0,8968	0,8699
REG-4-500	0,8827	0,8738	0,8314	0,9144	0,8873
REG-14-100	0,8439	0,6613	0,8274	0,6162	0,8380
REG-14-500	0,8772	0,8044	0,8694	0,6683	0,8714



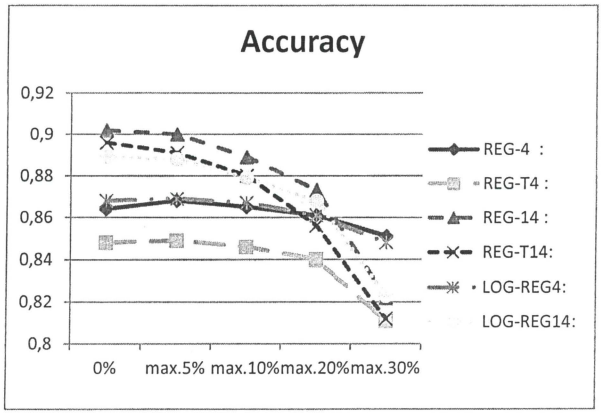
Rysunek 13. Porównanie *Precyzji* liniowych klasyfikatorów regresyjnych dla zmiennych parametrów modelu dla różnych liczności zbiorów uczących

Podobne efekty zauważamy analizując wpływ liczności zbioru uczącego na własności klasyfikatorów wykorzystujących regresję logistyczną. Należy również zauważyć, że dla niektórych charakterystyk jakości klasyfikatorów, np. *Czułości*, wpływ liczności zbioru uczącego jest mniej widoczny niż to ma miejsce w przykładach pokazanych na Rysunkach 12 – 13.

Trzeci z przeprowadzonych eksperymentów symulacyjnych miał odpowiedzieć na pytanie jak zmienia się efektywność klasyfikacji gdy parametry modelu generującego klasyfikowane obiekty ulegają losowym zmianom. W tym eksperymencie dopuszczaliśmy przypadek, że zmianie mogą jednocześnie ulec wartości oczekiwane wszystkich zmiennych objaśniających. Przedstawione w Tabelach 15 – 16 oraz na Rysunkach 14 – 15 wyniki dotyczą przypadku gdy wartości oczekiwane zmiennych objaśniających różnią się o maksimum *a%* od wartości oczekiwanych przyjętych w procesie generacji danych zbioru uczącego, przy czym zmiany te mogą mieć charakter zarówno wzrostu i zmniejszenia wartości parametru. Na Rysunku 14 porównano wartości *Dokładności* rozpatrywanych klasyfikatorów, a na Rysunku 15 analizowano wskaźnik *F1*.

Tabela 15. Porównanie *Dokładności* klasyfikatorów regresyjnych dla zmiennych parametrów modelu dla różnych poziomów zmienności

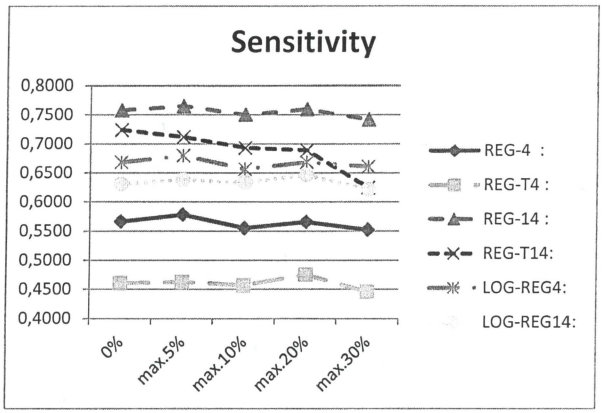
	0%	max.5%	max.10%	max.20%	max.30%
REG-4 :	0,864	0,868	0,865	0,861	0,851
REG-T4 :	0,848	0,849	0,846	0,84	0,811
REG-14 :	0,902	0,9	0,889	0,873	0,822
REG-T14:	0,896	0,891	0,88	0,856	0,812
LOG-REG4:	0,868	0,869	0,867	0,861	0,848
LOG-REG14:	0,889	0,888	0,879	0,868	0,823



Rysunek 14. Porównanie *Dokładności* klasyfikatorów regresyjnych dla zmiennych parametrów modelu dla różnych poziomów zmienności

Tabela 16. Porównanie *Czułości* klasyfikatorów regresyjnych dla zmiennych parametrów modelu dla różnych poziomów zmienności

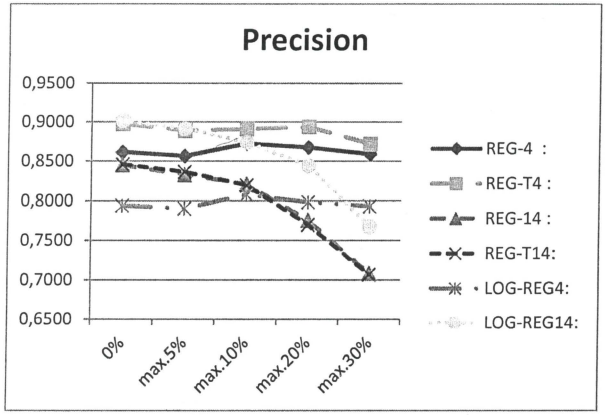
	0%	max.5%	max.10%	max.20%	max.30%
REG-4 :	0,5659	0,5778	0,5549	0,5653	0,5520
REG-T4 :	0,4606	0,4625	0,4568	0,4752	0,4457
REG-14 :	0,7577	0,7647	0,7500	0,7593	0,7414
REG-T14:	0,7236	0,7115	0,6928	0,6885	0,6240
LOG-REG4:	0,6679	0,6796	0,6561	0,6680	0,6603
LOG-REG14:	0,6307	0,6381	0,6333	0,6462	0,6217



Rysunek 15. Porównanie *Czułości* klasyfikatorów regresyjnych dla zmiennych parametrów modelu dla różnych poziomów zmienności

Tabela 17. Porównanie *Precyzji* klasyfikatorów regresyjnych dla zmiennych parametrów modelu dla różnych poziomów zmienności

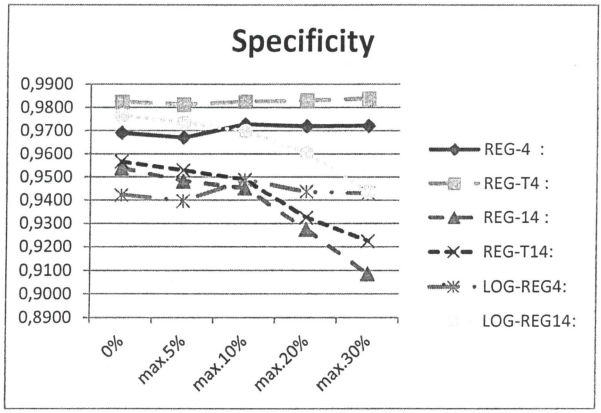
	0%	max.5%	max.10%	max.20%	max.30%
REG-4 :	0,8614	0,8564	0,8722	0,8673	0,8589
REG-T4 :	0,8974	0,8889	0,8911	0,8938	0,8719
REG-14 :	0,8448	0,8326	0,8208	0,7754	0,7083
REG-T14:	0,8458	0,8363	0,8194	0,7700	0,7069
LOG-REG4:	0,7940	0,7901	0,8076	0,7982	0,7931
LOG-REG14:	0,9016	0,8912	0,8737	0,8441	0,7677



Rysunek 16. Porównanie *Precyzji* klasyfikatorów regresyjnych dla zmiennych parametrów modelu dla różnych poziomów zmienności

Tabela 18. Porównanie *Specyficzności* klasyfikatorów regresyjnych dla zmiennych parametrów modelu dla różnych poziomów zmienności

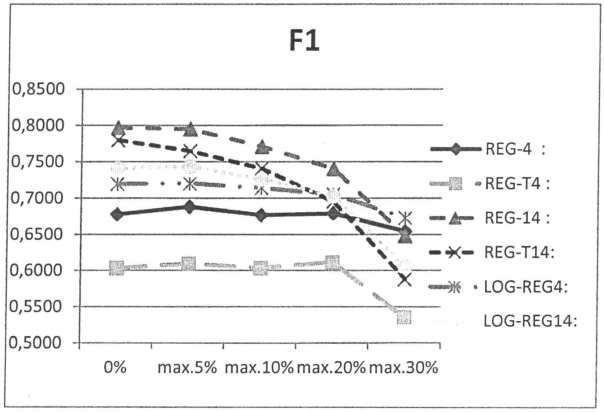
	0%	max.5%	max.10%	max.20%	max.30%
REG-4 :	0,9691	0,9671	0,9728	0,9719	0,9722
REG-T4 :	0,9824	0,9814	0,9826	0,9831	0,9839
REG-14 :	0,9538	0,9483	0,9451	0,9274	0,9086
REG-T14:	0,9566	0,9529	0,9489	0,9325	0,9225
LOG-REG4:	0,9423	0,9396	0,9489	0,9437	0,9430
LOG-REG14:	0,9767	0,9735	0,9698	0,9603	0,9436



Rysunek 17. Porównanie *Specyficzności* klasyfikatorów regresyjnych dla zmiennych parametrów modelu dla różnych poziomów zmienności

Tabela 19. Porównanie *wskaźnika F1* klasyfikatorów regresyjnych dla zmiennych parametrów modelu dla różnych poziomów zmienności

F1	0%	max.5%	max.10%	max.20%	max.30%
REG-4 :	0,6776	0,6883	0,6768	0,6789	0,6542
REG-T4 :	0,6033	0,6094	0,6034	0,6114	0,5356
REG-14 :	0,7966	0,7950	0,7708	0,7402	0,6477
REG-T14:	0,7794	0,7647	0,7407	0,6954	0,5881
LOG-REG4:	0,7196	0,7200	0,7139	0,7051	0,6728
LOG-REG14:	0,7407	0,7432	0,7260	0,7027	0,6055



Rysunek 18. Porównanie *wskaźnika F1* klasyfikatorów regresyjnych dla zmiennych parametrów modelu dla różnych poziomów zmienności

Analiza wyników eksperymentu wskazuje na zdecydowanie większą stabilność klasyfikatorów opartych na prostej regresji binarnej oraz prostej regresji logistycznej. W przypadku dużej zmienności parametrów (max. 30%) klasyfikator wykorzystujący prostą regresję binarną czterech zmiennych odznacza się również największą efektywnością. Jednakże w przypadku małych zmian wartości parametrów modelu najlepszymi pozostają klasyfikatory wykorzystujące kwadratową funkcję wiążącą, zarówno w przypadku regresji liniowej jak i regresji logistycznej.

4. Podsumowanie

W pracy porównano jakość klasyfikatorów zbudowanych na modelach regresyjnych opartych na prostej regresji liniowej oraz na regresji kwadratowej powierzchni odpowiedzi. Analizowano również jakość klasyfikatorów opartych na regresji logistycznej z liniową oraz kwadratową funkcją wiążącą. Porównania dokonano metodami symulacyjnymi dla przypadku gdy zmienne objaśniane są silnie zależne i są opisane rozkładami istotnie różniącymi się od rozkładu normalnego. Dodatkowo założono silnie nieliniową, a nawet nieróżniczkowalną, zależność pomiędzy zmiennymi modelu. Z przeprowadzonych badań wynika, że klasyfikatory wykorzystujące modele regresji binarnej są dokładniejsze od modeli wykorzystujących wyniki obserwacji zmiennych o wartościach rzeczywistych. Wykazują się one dodatkowo większą odpornością na zmiany parametrów modelu generacji danych. W warunkach w miarę stabilnych najlepszą jakością odznaczał się klasyfikator zbudowany na binarnej regresji kwadratowej powierzchni odpowiedzi. Jednakże w przypadku dużych zaburzeń modelu najlepszym okazał się klasyfikator wykorzystujący prostą regresję binarną.

Literatura

[1] Buja A., Stuetzle W., Shen Y. (2005, Nov. 3), *Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications* [Online]. Dostępne: *www-stat.wharton.upenn.edu/~buja.*

[2] Draper N.R., Smith H., *Analiza regresji stosowana*, PWN, Warszawa, 1973.

[3] Hastie T., Tibshirani R., Friedman J.: *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (2nd Ed.), Springer, New York, 2009.

[4] Hryniewicz O., SPC of processes with predicted data - application of the data mining methodology. W: *Frontiers in Statistical Quality Control 11*, S.Knoth, W.Schmid (Eds.), Physica Verlag, Heidelberg, 2015 (w druku).

[5] Hryniewicz O., Process inspection by attributes using predicted data, W: *Challenges in Computational Statistics and Data Mining*, J.Mielniczuk, S.Matwin (Eds.), Springer, (zgłoszone do publikacji).

[6] Hryniewicz O., Karpiński J. (2014), Prediction of reliability – pitfalls of using Pearson's correlation. Eksploatacja i Niezawodność - Maintenance and Reliability, 16, 472-483.

[7] Koronacki J. Ćwik J., *Statystyczne systemy uczące się*, WNT, Warszawa 2005.

[8] Nelsen R.B., *An Introduction to Copulas. 2nd edn.* Springer, New York 2006.

ROBUSTNES OF REGRESSION-BASED BINARY CLASSIFIERS AGAINST DEPARTURES FROM BASIC ASSUMPTIONS

Summary

The applicability of simple binary classifiers based on linear regression and quadratic response surface regression was discussed when the dependence between input and output variables is complex and nonlinear, and the model is different from the normal one. Cases when training sets (used for the construction of a classifier) and test sets (or data used in real applications) are generated by different mechanisms have been also considered. The results of simulation experiments suggest that binary classifiers based on the quadratic response surface regression are better than classifiers based on the simple linear regression when both training and test sets are generated by the same mechanism. However, classifiers based on the linear regression are more robust against a change of the mechanism that generates the data in real applications.

Key words: binary classifiers, linear regression, quadratic response surface regression, copulas

OLGIERD HRYNIEWICZ
Wyższa Szkoła informatyki Stosowanej i Zarządzania w Warszawie,
ul. Newelska 6, 01-447-Warszawa
hryniewi@ibspan.waw.pl
JANUSZ KARPIŃSKI
Instytut Badań Systemowych PAN
ul. Newelska 6, 01-447-Warszawa
ANNA OLWERT
Instytut Badań Systemowych PAN
ul. Newelska 6, 01-447-Warszawa