PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176
Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Julita Stańczuk, Patrycja Trojczak-Golonka
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

ANALIZA JAKOŚCI KLASYFIKACJI OBIEKTÓW Z NIEKOMPLETNYMI DANYMI Z WYKORZYSTANIEM SIECI NEURONOWYCH

Streszczenie: Celem artykułu jest przedstawienie znaczenia informacji opisowej dla klasyfikacji przedsiębiorstw notowanych na GPW w Warszawie, a dokładniej możliwości wystąpienia braków, danych zaszumionych czy celowej redukcji liczby zmiennych. Istotne jest to, w jaki sposób pogorszenie tej jakości wpływa na efektywność klasyfikacji, a więc przede wszystkim na liczbę przedsiębiorstw poprawnie zaklasyfikowanych do poszczególnych grup (z wykorzystaniem ratingu). Próbę badawczą tworzą przedsiębiorstwa notowane na GPW, dane natomiast pochodzą z ich sprawozdań finansowych. W badaniu wykorzystano sieci neuronowe umożliwiające m.in. klasyfikację obiektów. Posłużono się wcześniejszymi badaniami do porównania otrzymanych wyników.

Słowa kluczowe: klasyfikacja, sieci neuronowe, braki danych, szum.

1. Wstęp

Celem artykułu jest zaprezentowanie wpływu występowania braków oraz szumów informacyjnych w wektorach wejściowych na jakość klasyfikacji podmiotów gospodarczych z wykorzystaniem sieci neuronowej, a także metod radzenia sobie z tymi problemami.

Próbę badawczą stanowią przedsiębiorstwa notowane na GPW w Warszawie. Obiekty do badania wybrano metodą łatwości dostępu, co sprawia, że nie jest ona reprezentatywna. Selekcja zmiennych diagnostycznych, zgodnie z wcześniej przeprowadzonymi eksperymentami, oparta została na analizie wrażliwości [Stańczuk, Trojczak-Golonka 2010]. Dzięki temu dobrano 14 zmiennych, którymi są wskaźniki finansowe tych przedsiębiorstw, a dane do nich wygenerowano z ich sprawozdań finansowych. Zakres czasowy stanowią lata 2006-2007.

Zmienne charakteryzujące poszczególne obiekty to wskaźniki finansowe z zakresu rentowności, obrotowości i płynności. Do klasyfikacji podmiotów natomiast wykorzystano rating pięciostanowy, dzięki któremu badane jednostki można po-

dzielić na klasy od pierwszej, obejmującej spółki z najlepszym standingiem finansowym, do klasy piątej, w której znajdują się spółki o najgorszej kondycji.

2. Procedura badań

Do uczenia wykorzystano sieć jednokierunkową wielowarstwową (ang. Multilayer Perceptron − MLP). Trenowana była ona z nauczycielem ze względu na znane wartości ratingu dla każdego przedsiębiorstwa. Wykorzystano do tego celu Pakiet STATISTICA v. 9. Próbę badawczą przy każdym uczeniu sieci dzielono na trenującą i testującą w stosunku procentowym 80 : 20. Sieć ma 14 neuronów w warstwie wejściowej, jedną warstwę ukrytą z eksperymentalnie dobieraną liczbą neuronów ukrytych w zakresie od 1 do 50 (opcja wykorzystana w Automatycznym Projektancie Sieci w programie STATISTICA) oraz 5 neuronów wyjściowych.

3. Teoretyczny aspekt podejmowanej problematyki

Podczas analizy danych pochodzących z rzeczywistych obiektów, opisujących skomplikowane procesy w nich zachodzące, można napotkać liczne trudności, które często związane są z niepełnowartościowością danych. Mogą one być spowodowane niedokładnością albo błędami pomiaru, szumem informacyjnym, a także brakiem niektórych informacji.

Istnieje wiele powodów występowania niekompletnych danych pomiarowych, a w tym przypadku danych finansowych w sprawozdaniach przedsiębiorstw. Mogą to być przede wszystkim zaniedbania, zmiana zestawu zmiennych podczas procesu gromadzenia danych czy niejednorodne źródło ich pochodzenia.

O problemie brakujących danych mowa jest wtedy, gdy niektóre obiekty nie są opisane na całym zbiorze zmiennych, a więc w zgromadzonych danych brakuje niektórych wartości. Jednakże taka sytuacja nie powinna uniemożliwiać próby wnioskowania w oparciu o istniejące dane. Jeżeli chodzi o możliwości radzenia sobie z brakami w danych wejściowych, to przede wszystkim można wymienić następujące podejścia:

- Usuwanie obiektów lub zmiennych, w których braki te zostały zidentyfikowane. W badaniach metoda ta nie zostanie wybrana ze względu na cel artykułu. Wykorzystywana jest wtedy, gdy predykcja braków w zmiennych mało istotnych może powodować pogorszenie jakości klasyfikowania (dotyczy usuwania zmiennych). Usuwanie całych obiektów stosowane jest w sytuacji, gdy mamy do czynienia z wystarczająco obszerną bazą danych, w innych przypadkach jednak często usunięcie obiektów z brakami powoduje zaburzenia w możliwości poprawnego wnioskowania.
- Ignorowanie brakujących danych, a tym samym dostosowanie metody klasyfikacji obiektów do faktu występowania braków w danych. Wiele klasyfikatorów

radzi sobie z tym problemem, jednak specyfika sieci neuronowych wykorzystywanych w badaniach (perceptron wielowarstwowy) powoduje, że do badania niezbędny jest komplet zmiennych wejściowych.

- Uzupełnianie braków. Ten wariant zostanie wykorzystany w badaniach w niniejszym artykule.

Ostatnia metoda, szeroko omówiona w literaturze, bazuje na uzupełnianiu braków za pomocą pewnej wartości „wyliczonej" na podstawie mniej lub bardziej wyrafinowanego kryterium. Główną jej zaletą jest to, że umożliwia ona wykorzystanie sieci MLP do klasyfikacji po usunięciu braków, co z danymi niekompletnymi nie byłoby możliwe.

„Rekonstrukcji" zbioru danych, w którym znajdują się braki, można dokonać za pomocą następujących metod (wszystkie należą do procedur przetwarzania wstępnego, przed użyciem klasyfikatora):

- Uzupełnianie globalne obliczoną wartością dla wszystkich danych danej zmiennej. Najczęściej wykorzystywane są takie statystyki, jak średnia, mediana czy dominanta. Sposób ten należy do metod uzupełniania „bez nauczyciela". Warto dodać, że aby móc przykładowo zastosować średnią do usuwania braków, musi być spełnionych kilka warunków, m.in. rozkład normalny zmiennych, co zostało w analizowanym przykładzie sprawdzone.
- Uzupełnianie lokalne względem decyzji oraz zmiennej. Metoda możliwa do wykorzystania pod warunkiem posiadania informacji na temat klasyfikacji obiektów do klas decyzyjnych – „z nauczycielem". Po podzieleniu obiektów na klasy, wykorzystuje się w ich obrębie te same statystyki co w uzupełnianiu globalnym (np. średnia dla danego wskaźnika w obrębie jednej z klas).
- Uzupełnianie metodą *k*-najbliższych sąsiadów (*k-nearest neighbours*, k-nn) – bazowanie na założeniu o podobieństwie obiektów. Wartość zmiennej wyjściowej (zależnej) dla nowego punktu (w tym przypadku dla braku w danych) oceniana jest na podstawie zestawu *k* „przykładowych" punktów. Metoda *k*-najbliższych sąsiadów poszukuje tych *k* „przykładów" w najbliższym sąsiedztwie nowego punktu.

Problem szumu występującego w danych wejściowych, dla których wykorzystywane są sieci neuronowe, nie jest problemem krytycznym. Odpowiednio nauczona sieć neuronowa powinna dość skutecznie „odfiltrować" szum, dlatego też powszechnie uważa się, że sieć neuronowa jest odporna na szumy. Warto jednak eksperymentalnie sprawdzić, czy istnieje pewna granica, powyżej której nawet taki sposób klasyfikacji obarczony jest większym błędem ze względu na te szumy [Duch i in. 2000].

4. Analiza wpływu sposobu uzupełniania braków na jakość klasyfikacji

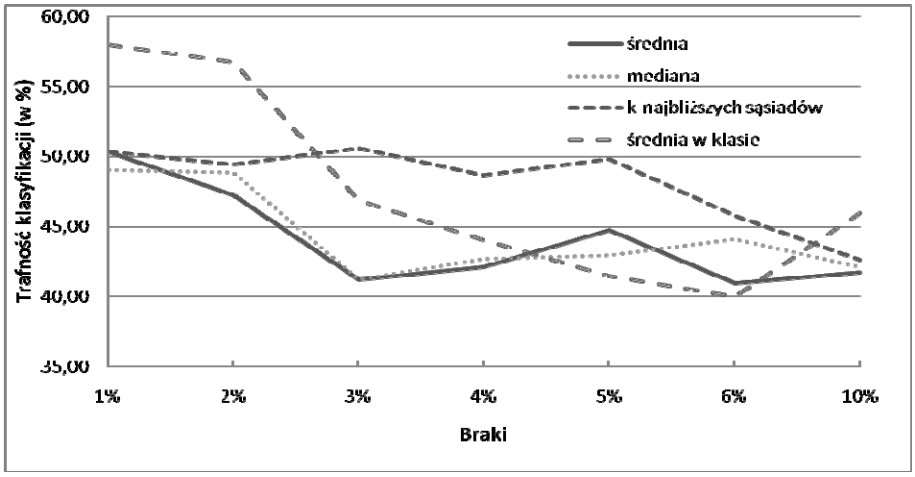
Trafność klasyfikacji, będąca ilorazem poprawnie sklasyfikowanych obiektów do wszystkich obiektów, posłużyła do oceny jakości klasyfikacji z wykorzystaniem wybranej metody uzupełniania braków.

W przypadku testowania sieci, gdzie zbiór danych jest kompletny, trafność wyniosła 53%. Wartość ta będzie stanowiła ważny punkt odniesienia dla otrzymanych wyników.

W celu wyeliminowania braków posłużono się następującymi metodami:

– uzupełnienie braków średnią dla danej zmiennej;
– uzupełnienie danych niepełnych medianą dla danej zmiennej;
– uzupełnienie z wykorzystaniem metody *k*-najbliższych sąsiadów. W metodzie tej jako optymalne, sprawdzone eksperymentalnie, przyjęto $k = 20$;
– uzupełnienie braków średnią dla danej zmiennej w klasie.

Braki w danych generowane były losowo w zakresie 1-10%, co przykładowo oznacza, że dla 1% braków wylosowano 40 atrybutów, które zostały usunięte. Wyniki przedstawiono na rys. 1.



Rys. 1. Trafność klasyfikacji przy losowych brakach w danych uzupełnianych różnymi metodami

Źródło: opracowanie własne.

Jak można zauważyć, uzupełnianie braków danych oparte na algorytmie *k*-najbliższych sąsiadów osiąga lepsze wyniki niż średnia czy mediana (globalne). Trafność klasyfikacji dla braków w zakresie 1-5% oscyluje między 49 a 51%. Jakość spada na poziomie 6% do 45%, natomiast przy 10% braków klasyfikacja jest

trafna już jedynie w 45%. Dla średniej i mediany (wyniki niewiele się różnią, o około 1%, jedynie w przypadku 6-procentowych niekompletnych danych jest to ponad 3%) trafność ta waha się w przedziale 40-50%, początkowo maleje wraz ze wzrostem danych niekompletnych, nieco wzrasta przy 5%, a następnie wraca do poziomu około 40%.

Jeśli natomiast porównać te wyniki z próbą likwidacji braków za pomocą średniej dla każdego wskaźnika w obrębie danej klasy, to przy niewielkich brakach, rzędu 1-2%, metoda ta jest zdecydowanie skuteczniejsza, nawet w porównaniu z algorytmem *k*-najbliższych sąsiadów. W przypadku 3% braków skuteczność klasyfikacji zbioru testującego zawiera się między metodami k-nn a średnią/medianą globalną. W pozostałych przypadkach jej trafność to plus/minus 2% w stosunku do metod uzupełniania globalnego średnią i medianą.

Jedynie dla 1% i 2% braków trafność klasyfikacji przy ich uzupełnianiu średnią lokalną uzyskano wyniki lepsze niż w przypadku sieci, w której posiadano komplet danych. W pozostałych przypadkach są to wartości niższe niż 53%. Jest to zrozumiale, ponieważ ten sposób uzupełniania powoduje powstawanie obiektów bardziej „charakterystycznych" dla danej klasy, co wynika z uśrednienia wartości atrybutów.

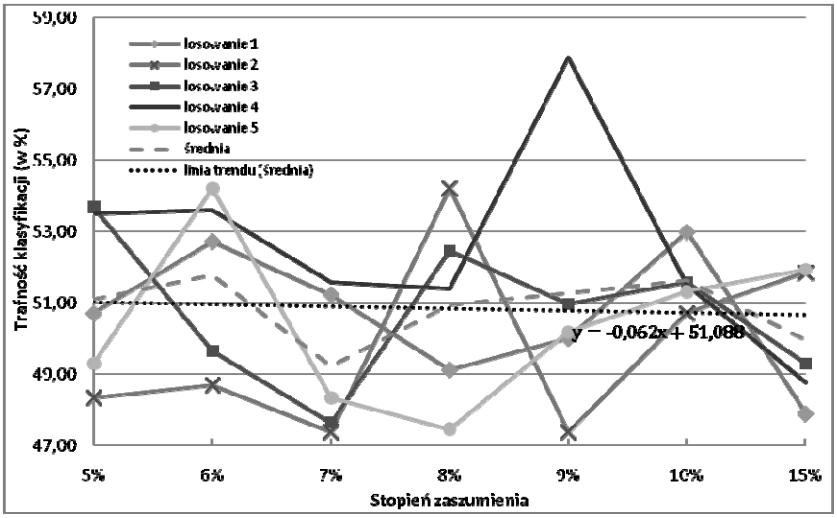
Podsumowując – wraz ze wzrostem liczby brakujących danych jakość klasyfikacji spada. Jedynie dla niewielkich ubytków rzędu 1-2% można uzyskać dobre wyniki pod warunkiem wykorzystania metody uzupełniania względem średniej w klasie. Jeśli zaś liczba braków przekracza 2%, należy korzystać z metody *k*-najbliższych sąsiadów.

5. Wpływ występowania szumów na klasyfikację obiektów

Kolejnym etapem badania jest sprawdzenie, jaki wpływ na jakość klasyfikacji ma zaszumienie danych. Zmienne oraz obiekty wybrano losowo, a następnie celowo zmieniono wartości atrybutów. Zrobiono to według algorytmu, który zakładał zmianę danych dla określonej, wcześniej zadeklarowanej, liczby atrybutów (oznaczone jako stopień zaszumienia, np. dla 10% jest to 400 atrybutów). W praktyce oznaczało to zwiększenie/zmniejszenie wartości konkretnej zmiennej o od 1 do 10% (również wartość oraz znak zmiany był dobierany losowo).

Taka procedura miała spowodować, że dane zostaną zaszumione w sposób losowy, zarówno pod względem wyboru zaszumionych atrybutów, jak i zmiany wartości wskaźnika. W celu uzyskania lepszego obrazu badanego zjawiska eksperyment został przeprowadzony wielokrotnie. Wybrane wyniki przedstawiono na rys. 2. Wynika z niego, że trafność klasyfikacji w znacznym stopniu zależy od losowego zaszumienia próbek w zbiorze wejściowym. Średnia trafność oscyluje w granicach 49-51%. Jest to niewiele mniej niż klasyfikacja wzorcowej próbki (53%). Wyznaczona linia trendu wskazuje na nieznaczny spadek jakości klasyfikacji w obrębie

badanych obiektów. Zauważalne jest również to, że wraz ze wzrostem zaszumienia rozrzut wyników jest coraz mniejszy.



Rys. 2. Trafność klasyfikacji dla sieci neuronowej z losowo zaszumionymi danymi

Źródło: opracowanie własne.

Minimalna trafność klasyfikacji to około 47% dla danych, w których znajdowało się 7 i 9% próbek zaszumionych. Maksymalna zaś to prawie 58% dla jednego z losowań danych o 9% stopniu zaszumienia. Jednak w pozostałych przypadkach jest to około 51-53%.

6. Podsumowanie

Na podstawie przeprowadzonych badań można wyciągnąć wnioski dotyczące postawionego problemu badawczego:

- Na jakość klasyfikacji w znacznym stopniu wpływa sposób uzupełniania braków występujących w danych wejściowych, będący jedną z metod radzenia sobie z nimi. W przypadku niewielkich ubytków, rzędu 1-2%, jakość klasyfikacji przy wykorzystaniu dla braków średniej dla danej zmiennej z klasy jest porównywalna lub czasami wyższa od wzorcowej (dla próbki z kompletem danych). Jeśli jednak zwiększa się niekompletność danych, wtedy lepsza okazuje się metoda *k*-najbliższych sąsiadów, przy czym jakość początkowo nieznacznie spada, po czym około 10% braków znacznie wpływa na pogorszenie trafności.
- Losowy dobór zmiennych oraz przypadków, których wartości zostały sztucznie zmienione według opisanego algorytmu, sprawia, że trafność klasyfikacji nie-

znacznie spada. Jakość klasyfikacji ulega również silnym wahaniom w zależności od losowania. Jednak średnia trafność to około 51%, a więc można potwierdzić powszechne twierdzenie o odporności sieci na szumy.

Literatura

Duch W., Korbicz J., Rutkowski L., Tadeusiewicz R. (red.), *Biocybernetyka i inżynieria biomedyczna. Sieci neuronowe*, Akademicka Oficyna Wydawnicza ELIT, Warszawa 2000.

Stańczuk J., Trojczak-Golonka P., *Wpływ zróżnicowania zbiorów atrybutów i procesu walidacji na efektywność klasyfikacji przedsiębiorstw przy wykorzystaniu sieci neuronowych*, [w:] K. Jajuga, M. Walesiak (red.), *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 17, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2010.

ANALYSIS OF THE QUALITY OF ECONOMIC OBJECT CLASSIFICATION WITH INCOMPLETE DATA USING NEURAL NETWORKS

Summary: The aim of the article is presenting the importance of descriptive information for the classification of companies listed on the Warsaw Stock Exchange, and more the possibility of missing data, corrupted by noise variables or deliberate reduction the number of variables. It is essential how the deterioration in the quality of input data affects the efficiency of classification – the number of companies correctly classified into particular groups (classification using rating). Companies listed on the WSE are creating research material, while the data are derived from their accounts. In the article artificial neural networks are used, which allow inter alia objects classification. Previous analysis are used to compare the results obtained with the classification using the complete information about the companies.