Wydział Ekonomii i Zarządzania i Turystyki
UNIWERSYTETU EKONOMICZNEGO
WE WROCŁAWIU
PISMO WPŁYNĘŁO
15.12.2014

Dr hab. Andrzej Sokołowski, Prof. UEK
Zakład Statystyki

Recenzja pracy doktorskiej mgr Karoliny Moniki Bartos pt. „Sieci neuronowe w badaniach zachowań konsumentów"

Hipoteza główna jest chyba zbyt ogólna. Prowokacyjnie można zadać pytanie, czy hipotezę taką można zweryfikować bez pisania tej pracy doktorskiej. Wydaje mi się, że do stwierdzenia przydatności sieci neuronowych do segmentacji konsumentów, analizy odejścia klientów oraz analizy koszykowej wystarczy wskazać po jednym zastosowaniu wziętym z literatury. A takie przykłady zawiera właśnie Tabela 2.1 w pracy.

Jednak tytuł pracy bardzo dobrze oddaje jej zawartość. Mocną strona rozprawy jest jej struktura, świetne wyważenie proporcji rozdziałów i na ogół dobry język prezentacji całości. Autorka ominęła niebezpieczeństwa zbytniego wdawania się w szczegóły teoretyczne sieci neuronowych, a jednocześnie nie pominęła żadnego ważnego aspektu budowy sieci. Może trochę zabrakło mi uwag krytycznych dotyczących sytuacji, gdy sieć daje słabe prognozy lub źle identyfikuje obiekty. Wtedy w zasadzie nie dostarcza podpowiedzi dlaczego tak się dzieje.

W trzech rozdziałach przedstawiono trzy różne przykłady zastosowania sieci neuronowych do badania zachowań konsumentów. Dotyczą one segmentacji klientów, analizy ryzyka odejścia klienta oraz analizy koszykowej – trzech ważnych obszarów, co pośrednio miało służyć do wskazania na pewną uniwersalność sieci neuronowych. Te przykłady zastosowań przedstawione są bardzo dobrze. Opisano problem, scharakteryzowano dane statystyczne, charakter stosowanej sieci neuronowej, a wreszcie przedstawiono wyniki. Sposób prezentacji jest jasny i zrozumiały, a w toku przedstawionych analiz nie znajduję żadnych błędów. Można by co najwyżej w niektórych miejscach nieco poszerzyć analizy o elementy spoza zakresu sieci neuronowych, co wskazuję w uwagach szczegółowych. To jednak nie ma wpływu na wartość pracy.

Autorka sprawnie zastosowała metody, które wymagają jednak pewnego doświadczenia badacza i wbrew pozorom nie są procedurami „automatycznymi". Wskazane poniżej usterki dotyczą głównie warstwy redakcyjnej pracy oraz pewnych sformułowań szczegółowych.

Stwierdzam, że praca spełnia ustawowe wymogi stawiane rozprawom doktorskim i wnoszę o dopuszczenie Autorki do publicznej obrony.

Uwagi szczegółowe

Strona (s) 7, wiersz (w) 5, od góry (g) – Niezręczne jest „opieranie się" na przedstawionych rozważaniach teoretycznych

s.9, Rys.1.1 – Trochę zabrakło mi informacji o charakterze sygnałów przesyłanych w ramach neuronów. Czy taki sygnał jest tylko binarny (jest / nie ma), czy ważne jest jego natężenie, czy ważny jest czas trwania, czy ma jakieś charakterystyki harmoniczne?

s.11, wzór (1.1.) – We wzorze tym subskrypt *i* oznacza numer wejścia, natomiast w pozostałych wzorach podawanych w pracy jest to numer wyjścia

s.14, w.15, g – *razy* zamiast *raz*

s.15, w.4, od dołu (d); s.33, w.3, d; s.67, w.2, g– *liczby* zamiast *ilości*

s.16, w.11, d – Zdanie rozpoczynające się od „Rozpatrując ..." ma tzw. składnię kabaretową (*Będąc młodą lekarką wszedł raz do mej przychodni pacjent*) charakteryzującą się zmianą podmiotu w trakcie zdania

s.20, w.15, d – Uwaga do zdania „Dane ze zbioru ...": Precyzyjnie, to nie dane są skorelowane tylko cechy (zmienne)

s.20, w.9, d – *obok* zamiast *koło*

s.20, w.5, d – „profesor" przez jedno s

s.21 – Brak precyzji w opisie symboli. Bezpośrednio pod rysunkiem jest napisane, że $x_1,...,x_n$ to neurony, a w drugim wierszu od dołu, że są to wartości. Jest tu jednak taka różnica jak pomiędzy zmienną, a jej realizacją. Lepiej byłoby na oznaczenie neuronów użyć wielkich X-sów. Nie opisano znaczenia subskryptów, choć oczywiście łatwo domyśleć się, że *i* to numer neuronu warstwy wyjściowej, a *j* to numer neuronu warstwy wejściowej. Jakoś wydaje mi się, że notacja odwrotna byłby bardziej naturalna. W zapisie pod rysunkiem wagi mają dwa subskrypty, a w drugim wierszu od dołu tylko jeden. Warto było oznaczyć wektory czcionką pogrubioną.

s.22, wzór (1.5) – Nieprecyzyjny zapis wzoru. Po prawej stronie wagi w_{ij} (dlaczego nie w_{ji} skoro ruch jest od wejścia do wyjścia a nie na odwrót) mają również subskrypt *i* „po którym" się nie sumuje, wobec tego wynik po lewej stronie też musi mieć ten subskrypt. Odejmowanie i sumowanie po prawej stronie sugeruje, że wszystkie x-sy i wszystkie wagi muszą być wyrażone w tych samych jednostkach. Brak komentarza jak to jest zapewniane. Te same uwagi dotyczą wzoru (1.6).

s.22 – Pytanie prowokacyjne. Czy nie najprościej byłoby podstawić jako wagi wartości x-sów, bo wtedy wagi najlepiej „pasują" do wejść i wtedy odległość w (1.7), pomiędzy wektorem wzorca wejściowego, a wektorem wag byłaby najmniejsza (Patrz też zdanie: *Podczas procesu uczenia ...* na s.25). Znów szkoda, że wektorów nie oznaczono czcionką pogrubioną.

s.27 – W opisie wzorów (1.10) i (1.11) pomylono *numery* wektorów uczących (u) z *liczbą* tych wektorów (z). Dotyczy to również numerów i liczby wejść. Szkoda, że precyzyjnie nie

określono wzajemnych relacji takich pojęć jak: wejście, wyjście, cecha (zmienna), wartość cechy, obiekt, identyfikator przynależności obiektu.

s.30, w.15, d – To modne ostatnio „odkrywanie wiedzy z danych" to nic innego jak statystyka, której zadaniem jest od wieków „poszukiwanie prawidłowości w procesach masowych".

s.38, w.2, d – *ilości* zamiast *ilość*

s.40, w.8, d – Trudno uznać, że to dopiero rozwinięcie teorii Beckera zapoczątkowało odkrywanie związków między wiekiem konsumenta a kupowanymi przez niego produktami. Do tego wystarczył zwykły rozsądek. Zakupy zabawek, piłek, sztucznych szczęk, czy lasek są w sposób oczywisty powiązane z wiekiem konsumenta.

s.42, w.1, d – Nie sądzę, aby prawdziwe było zdanie *Nabywcy wybierają i kupują bowiem tylko te produkty, które najpełniej zaspokajają ich potrzeby*. Niewątpliwie Ferrari najpełniej zaspokoiło by rozliczne potrzeby wielu mężczyzn, ale przecież nie kupujemy tylko samochodów tej marki.

s.54, w.10 – Dlaczego do zakodowania płci nie wystarczy tylko jedna zmienna. A jeżeli używamy dwóch to co z problemem współliniowości (redundancji)?

s.61, w.1, g – Nieprawdziwe jest stwierdzenie o *Istnieniu nieskończonej ilości potrzeb człowieka* .... Jako ćwiczenie proponuję wymienienie 10000 własnych potrzeb.

s.63, Rys.3.2 – Wątpliwa jest krzywa Tornquista dla dóbr luksusowych. Trudno sobie wyobrazić, że przy braku ograniczeń finansowych, ktoś nabędzie 100 Mercedesów, 50 domów na Lazurowym Wybrzeżu, czy 30 jachtów (a bardziej przyziemnie – 500 flakoników różnych perfum).

s.66, w.3, d – Nie wyobrażam sobie takiego zagadnienia, w którym metoda k-średnich potrzebowałaby aż 200 iteracji to znalezienia ostatecznego podziału.

s.74, Tab.3.9 – Szkoda, że nie poddano testowaniu istotności różnic średnich udziałów w grupach, choćby przybliżonemu, przy pomocy jednoczynnikowej analizy wariancji.

s.75, ostatni akapit na stronie – Dla mnie nieprzekonywująca jest przewaga SOM nad metodą k-średnich w sensie możliwości oceny, które skupienia są podobne, a które się różnią bardziej. Takie informacje zawiera Tabela 3.2 i nieważne w jaki sposób wykonywano grupowanie.

s.81, pierwszy akapit – Nieco złośliwa opinia głosi, że zadaniem marketingu jest wciskanie ludziom rzeczy niepotrzebnych. Paradoksalnie potwierdza to duży procent bezrobotnych wśród absolwentów kierunku zarządzanie i marketing.

s.88, Tab. 4.4 – Wpływ terminu obrony na decyzje o kontynuowaniu studiów należało zweryfikować przy pomocy testu niezależności chi-kwadrat.

s.88, Rys.4.4, Rys.4.5 – Uważam, że wpływ średniej oceny na kontynuowanie studiów badany poprzez porównywanie rozkładów jest niewłaściwy. Liczba osób o średniej 3,6-4,0 jest po prostu duża. Można było badać rozkład tak/nie w zależności od przedziałów ocen (znów test niezależności chi-kwadrat). Wykres ramkowy 4.6. sugeruje brak wpływu średniej oceny na decyzję. Zapewne potwierdził by to test dla dwóch średnich, którego nie zastosowano.

s.90 – Chyba niepotrzebnie oddano ocenę wartości predykcyjnej cech automatowi *STATISTICI*. Nie zawsze podział zakresu zmienności zmiennej ciągłej na cztery przedziały (równe?) jest najbardziej sensowny. Co oznacza tajemnicza „ważność" statystyki chi-kwadrat na Rys.4.7? (Statystyka nie zna pojęcia „ważności statystyki testowej"). Chyba nie jest to wartość statystyki chi-kwadrat, bo wartości te są porównywalne tylko w przypadku tej samej liczby stopni swobody. Sensowny byłby iloraz empirycznej wartości statystyki przez wartość teoretyczną przy zadanym poziomie istotności.

s.93, Tab.4.8 – Szkoda, że przy opisie wyników klasyfikacji nie użyto popularnych określeń „czułość" (tu 0,87) i specyficzność (0,69). Należało policzyć też *dodatnią zdolność predykcyjną* (0,80) oraz *ujemną zdolność predykcyjną* (0,79)

s.98, w.5, g – Zła odmiana nazwiska Agnieszki Pasztyły

s.101, w.12-13, d – Na ogół statystycy zgadzają się z definicją Kołmogorowa według której prawdopodobieństwo to liczba niemianowana z przedziału [0,1] i nie może wynosić 80%.

s.101, w.4, d – Co to jest metoda *a priori*?

s.105, w.8, g – W jaki sposób odróżnia się zachowanie według reguły *jeżeli papierosy, to piwo*, od zachowania według reguły *jeżeli piwo, to papierosy*?

s.115, pierwsze zdanie pod Rys. 5.4 – Ma być *liczba* zamiast *ilość*

s.116, Tablica 5.18 – Braki liter w wyrazach *błyskawiczne* i *doładowanie*

s.127, w.4, d – Niezręcznie *opierano się na dostępnych danych*

s.129, s.12, g – Nie ma *metody doboru i eliminacji zmiennych pakietu STATISTICA*; jesteś jakaś metoda, którą zaimplementowano w programie *STATISTICA*