Prof. dr hab. inż. Jacek Koronacki
Instytut Podstaw Informatyki PAN
Jana Kazimierza 5, 01-248 Warszawa

Warszawa, 4 listopada 2013

Recenzja rozprawy doktorskiej mgr Diany Domańskiej
Prognozowanie eksploracyjne danych wielowymiarowych

Zakres tematyczny rozprawy

Tematykę pracy dobrze oddaje jej tytuł. Zadanie prognozowania odnosi nas natychmiast do obszaru analizy szeregów czasowych, w tym przypadku wielowymiarowych. Jest to obszar tyleż klasyczny, co bardzo ... nieklasyczny i nowoczesny – bogactwo danych biologicznych, ubezpieczeniowych, finansowych, przemysłowych i wielu innych wymusiło ogromny rozwój klasycznych metod analizy szeregów czasowych, w szczególności, ale nie tylko, objęcie przypadków z niejednorodną wariancją, nieliniowych oraz o zależności dalekiego zasięgu. Często chodzi tu o szeregi wprawdzie wielowymiarowe, ale indeksowane tylko przez czas, ponieważ zbierane w jednym miejscu. Często jednak chodzi też o dane przestrzenne, czyli zbierane w różnych punktach przestrzeni i tym samym także przestrzennie, a nie tylko przez czas, indeksowane. Właśnie takie dane są przedmiotem zainteresowania Doktorantki.

Chodzi mianowicie o dane dotyczące imisji różnego typu zanieczyszczeń powietrza, zbierane w różnych miejscach np. w Polsce. Statystyk gotów jest w takim przypadku przywołać natychmiast modele statystyczne dla danych przestrzennych, ale prawidłowa, wiarygodna estymacja parametrów takich modeli wymaga dokonywania pomiarów w dostatecznie wielu punktach przestrzeni. A z taką sytuacją nie mamy do czynienia w przypadku danych analizowanych przez mgr Domańską – dane pochodzą z zaledwie kilku stacji pomiarowych, do tego bardzo od siebie odległych.

Doktorantka musiała zatem spojrzeć na problem inaczej. Mogła albo powielić metody już stosowane, albo zaproponować własne podejście, i wybrała to drugie rozwiązanie. Nie mogąc powiązać danych z różnych punktów przestrzeni w jeden model przestrzennie indeksowany, wybrała podejście opierające się na poszukiwaniu w przeszłości przebiegów w odpowiednim sensie podobnych i na tej podstawie budowie prognoz, które nazwała (słusznie) prognozami eksploracyjnymi. Musiała zatem opracować metody pomiaru podobieństwa i na tej podstawie własne metody prognozy. Oparła się na zastąpienia ciągów liczbowych liczbami rozmytymi, postępując w sposób zarazem naturalny i oryginalny (autorski).

Na rozprawę składa się pięć rozdziałów oraz cztery dodatki. Dwa pierwsze rozdziały mają charakter wprowadzenia, w trzech następnych znajdujemy autorskie algorytmy dające podstawę do rozmytego grupowania i budowy algorytmów prognozy (rozdziały 3 i 4) oraz szeroki materiał eksperymentalny (rozdział 5). W dodatkach znajdujemy informacje dodatkowe, np. o typach i charakterze zanieczyszczeń powietrza oraz kolejne wyniki analiz eksperymentalnych. Wyniki zawarte w rozdziale 5 pozwalają mgr Domańskiej dokonać porównania jej podejścia z podejściami znanymi w literaturze, wyniki zawarte w dodatku D są oryginalne i ciekawe, ale nie ma w literaturze porównywalnych eksperymentów.

Na str. 7 pracy mgr Domańska pięknie podsumowała cele pracy oraz przedstawiła jej tezę.

Rozprawa liczy 149 stron. Na bibliografię składa się 126 pozycji. Doktorantka miała rację zamieszczając spis oznaczeń (minimalnie zbyt krótki), spis rysunków i tabel.

Uwagi ogólne

Rozprawę należy ocenić jako bardzo dobrze skomponowaną – całość wywodu jest logiczna i spójna. Jest to praca prezentująca stosunkowo szerokie, erudycyjne i nowoczesne spojrzenie na jej przedmiot. Ładunek oryginalności odpowiada wymaganiom, jakie stawia się rozprawom doktorskim. Pozytywnie uderza systematyczność wywodu – na str. 22 otrzymujemy jasne i zwarte opisanie metodologii tworzenia prognozy, ogólne sformułowanie problemu prognozy, zaś w dalszym ciągu rozprawy każdy kolejny krok postępowania ma swoją jasno opisaną motywację i metodologiczną podstawę. Każdy kolejny podrozdział jest logiczną konsekwencją podrozdziału go poprzedzającego.

Oryginalny wkład autorski jest duży i z pewnością, a nawet z pewnym naddatkiem, spełnia wymagania stawiane rozprawom doktorskim. Gdyby ów – jak go nazwałem – naddatek dotyczył materiału teoretycznego, związanego z podstawami metodologicznymi, a nie ogromu głębokiej, rzetelnej i szerokiej analizy wyników eksperymentalnych, rozprawa na pewno zasługiwałaby na wyróżnienie.

Mam tylko trzy ogólne uwagi o charakterze krytycznym lub dyskusyjnym.

Doktorantka zbyt mocno oparła swój ogólny opis metod prognozowania na skądinąd bardzo dobrej książce napisanej po redakcją J.S. Armstronga w roku 2001. Pod wieloma względami był to wybór słuszny, ponieważ pozwolił Jej dobrze opisać problem wyboru metody prognozowania w przypadku Ją interesującym. Ale pod pewnymi względami był to wybór zbyt ograniczający ogólną analizę metod prognozowania. Na przykład odwołanie się do „metod ekonometrycznych" jest odwołaniem się do metod popularnych wśród ekonometryków, a nie do całej gamy metod statystycznych. W rezultacie akapit odnoszący się do tych ostatnich jest zdecydowanie zbyt ubogi, oparty na literaturze klasycznej, wartościowej, ale pomijającej wszystko co nieklasyczne.

Wyniki analiz porównawczych z rozdziału 5, jak już napisałem godne pochwały z racji ich rzetelności, w tym oparciu się na porównywaniu wyników z miejsc możliwie podobnych, mają chyba wartość nie w pełni satysfakcjonującą. To prawda, że chcąc porównać swoje podejście z innymi, znanymi z literatury, naturalne było wykorzystanie także wyników znanych z literatury. Ale to skazywało Doktorantkę na porównywanie zanieczyszczeń np. w Zagrzebiu z zanieczyszczeniami w Złotym Potoku czy Wodzisławiu Śląskim. Czy naprawdę można w przypadku tak różnych typów miejsc zbierania danych sądzić, że takie porównania są w pełni uzasadnione? Czy nie należało zaimplementować różnych metod prognozy i zastosować ich do danych z tego samego miejsca?

W przypadku każdej z dokonanych analiz porównawczych mgr Domańska bardzo jasno i trafnie podsumowała otrzymane wyniki. Ale nigdy nie zadała sobie pytania, a zatem także nie udzieliła odpowiedzi na pytanie dlaczego w danym przypadku to ta a nie inna metoda wypadła najlepiej.

Uwagi szczegółowe

Uwagi te są niekiedy trywialne, a jeśli takie nie są, to dotyczą bardzo drobnych usterek redakcyjnych i nie mają wpływu na bardzo pozytywną ocenę całości rozprawy.

- s.2: *[...] możliwe stało się*, a nie *możliwie stało się*;
- s. 10: Autorka pisze: *[...] dwa typy szeregów nazywane są ciągłymi i dyskretnymi [...], pomimo tego że w obu przypadkach zmienna mierzona może być dyskretna lub ciągła. [...] Dla ciągłych szeregów czasowych zmienna obserwowana jest zmienną ciągła.* Stwierdzenia te są w oczywisty sposób sprzeczne.
- s. 19: Autorka pisze: *Ostatnią interpretacją wartości funkcji przynależności jest stopień niepewności. W tym przypadku wartość przynależności opisuje stopień wiarygodności [...].* Stopień wiarygodności to raczej 1 – stopień niepewności.
s. 20: Autorka pisze: *Liczby, które będą miały najwyższą wartość przynależności z dyskretnego zbioru można potraktować jako skupienie liczb wokół liczby rozmytej.* Czy określenie *wokół* jest najbardziej szczęśliwe?
- s. 22-23: współczynnik korelacji Pearsona (wzór (2.1)) to oczywiście to samo co korelacja jako standaryzowana kowariancja (wzór (2.2)). Dlaczego więc są potraktowane jako różne miary? Zresztą wszystkie dalsze opisy dotyczą tej samej korelacji, tyle że w różnych kontekstach. W tym punkcie dziwi brak dyskusji problemu prędkości zaniku korelacji między danymi z różnych chwil.
s. 24: Punkt 2.3 dotyczy danych jednowymiarowych. To nic złego, ale czytelnik powinien być o tym uprzedzony. Powinien także dowiedzieć się, przynajmniej w jednym zdaniu, jak zająć się przygotowaniem danych wielowymiarowych.
s. 29: Osądowy bootstrapping nie jest jasno opisany – ktoś kto nie wie, czym jest, z przedstawionego opisu nie dowie się tego. (Poza tym Autorka ma na myśli początek XXI, a nie XX, wieku.)
s. 40: Zdanie: *Celem predykcji jest przebieg czasowy zjawiska* warto chyba było uzupełnić słowami „w przyszłości".
Def. 3.2: Symbol d^{nm} nie jest zdefiniowany (choć jego znaczenie jest oczywiste).
s. 43: Zdanie zaczynające się od słów *Po pierwsze* jest niejasne; to nieprawda, że *suma prawdopodobieństw wszystkich zdarzeń ma wynosić 1.* Przykład z grą w Lotto jest niejasny.
s. 44: zmienna niezależna nie powinna być nazywana parametrem.
s. 46: Nie jest jasne skąd bierze się *wiele szeregów czasowych.* Nie jest napisane co oznacza optymalność parametrów α i β.
s. 49: Nie wiem, co miał na myśli zacytowany w motcie Paulo Coelho.
s. 53: Gdy piszemy horyzont czasowy T i za chwilę krok czasowy Δt, to sugerujemy, że niekoniecznie T jest liczbą naturalną, a to wynika z ciągu dalszego; innymi słowy, musimy mieć $\Delta t=1$.
s.53_8: Nie jest w tym miejscu jasne czym są sektory kierunkowe.

Konkluzja

Rozprawę oceniam jako bardzo wartościową i z pewnością spełniającą wymagania stawiane pracom doktorskim w dziedzinie nauk technicznych, w szczególności w dyscyplinie informatyka. Wnoszę o dopuszczenie mgr Diany Domańskiej do dalszych etapów przewodu doktorskiego.