

RECENZJA

rozprawy doktorskiej Pani mgr inż. Agnieszki SUCHWAŁKO
pt. *Zastosowanie analizy statystycznej do identyfikacji bakterii
na podstawie widm dyfrakcyjnych kolonii bakterii*

1. Zakres tematyczny rozprawy

Metody automatycznego rozpoznawania obiektów są szeroko stosowane w różnorodnych praktycznych zadaniach podejmowania decyzji. We współczesnym świecie nie ma praktycznie obszaru, w którym metody te byłyby nieobecne. Z jednej strony, wynika to z faktu, że wiele problemów z zakresu techniki, medycyny, biologii, czy zjawisk społecznych i ekonomicznych sprowadza się do zadania rozpoznawania o wspólnym opisie i metodach rozwiązania, niezależnie od dziedziny, z której pochodzą. Z drugiej strony, metodologia rozpoznawania, bardzo intensywnie rozwijana od paru dziesiątek lat, dysponuje wieloma metodami i technikami rozpoznawania, różniącymi się własnościami, paradygmatami budowy klasyfikatorów, czy schematem decyzyjnym, pokrywającymi tym samym szeroki zakres problemów praktycznych.

Recenzowana praca doktorska dotyczy właśnie praktycznego zastosowania wybranych metod formalnych w zadaniu automatycznego rozpoznawania gatunków bakterii. Właściwe rozpoznanie gatunku jest ważne dla diagnostyki medycznej (identyfikacja patogenów), a także stanowi istotny element kontroli skażenia mikrobiologicznego środowiska. Spośród wielu metod rozpoznawania bakterii stosowanych w laboratoriach mikrobiologicznych, doktorantka skupiła się na metodzie optycznej (opracowanej w ośrodku wrocławskim), w której podstawą rozpoznawania jest analiza widm dyfrakcyjnych kolonii bakterii. Jako cel badań przyjęto eksperymentalną analizę skuteczności wybranych klasyfikatorów, metod selekcji/redukcji cech i algorytmów wstępnego przetwarzania obrazów widm dyfrakcyjnych, ukierunkowaną na opracowanie maksymalnie efektywnej metody automatycznego rozpoznawania gatunków bakterii na podstawie widm dyfrakcyjnych Fresnela kolonii bakterii zarejestrowanych w dedykowanym układzie optycznym.

W kontekście powyższych uwag, tematykę rozprawy uważam za w pełni uzasadnioną i aktualną dla współczesnych prac z zakresu diagnostyki mikrobiologicznej oraz dla rozwoju metodologii automatycznego rozpoznawania, prowadzącą do wzbogacenia wiedzy o nowy, bardzo interesujący przykład aplikacyjny.

2. Zawartość rozprawy

Ponieważ podstawą do ubiegania się przez doktorantkę o stopień naukowy doktora n.t. stanowi rozwiązanie określonego problemu badawczego, którego poszczególne składowe zostały wcześniej opublikowane w 6 pracach (rozdział w książce, 2 prace w czasopiśmie z

listy JCR i 3 prace w materiałach konferencyjnych), dlatego zasadniczą treść rozprawy stanowią odbitki 6 prac przedstawione w formie załączników. Zostały one poprzedzone 4 rozdziałami zawierającymi syntetyczne ujęcie wyników oraz stanowiącymi przewodnik po szczegółach merytorycznych zawartych w publikacjach. Dodatkowo, na zawartość rozprawy składa się spis literatury obejmujący 110 pozycji oraz – co bardzo ważne – oświadczenia współautorów publikacji, jednoznacznie określające wkład doktorantki w rozwiązanie postawionych problemów badawczych.

Rozdział 1 zawiera szeroko nakreślony kontekst literaturowy w dziedzinie rozpoznawania bakterii, a więc w obszarze objętym tematyką rozprawy. Na szczególne podkreślenie zasługuje systematyczne podejście do analizy aktualnego stanu badań na świecie w omawianej dziedzinie oraz wyekspozowanie na tym tle osiągnięć ośrodka wrocławskiego z Katedry Inżynierii Biomedycznej kierowanej przez prof. H. Podbielską, którego bardzo aktywnym członkiem jest doktorantka. Przedstawiony *state-of-the-art* w pełni uzasadnia zakres tematyczny rozprawy oraz pokazuje znaczenie przeprowadzonych badań dla rozwoju metodologii diagnostyki mikrobiologicznej.

W rozdziale 2 w sposób jednoznaczny i wystarczająco szczegółowy przedstawiono cele badań, sformułowano tezy rozprawy oraz wymieniono najważniejsze uzyskane rezultaty.

Rozdział 3 zawiera w syntetycznym ujęciu wyniki zrealizowanych prac badawczych. Przedstawiono tu kolejno metodologię prowadzonych badań eksperymentalnych oraz uzyskane rezultaty w podziale na etapy prac badawczych. Podstawą prowadzonych badań eksperymentalnych były widma dyfrakcyjne kolonii bakterii zarejestrowane w specjalnie zaprojektowanym układzie optycznym. Aby je pozyskać, należało wykonać szereg czynności (przygotowanie odseparowanych próbek kolonii bakterii zgodnie z procedurami mikrobiologicznymi, oświetlenie wiązką laserową, rejestracja widm dyfrakcyjnych), które realizowane były we współpracy z Uniwersytetem Przyrodniczym we Wrocławiu. Na etapach wstępnych (od przygotowania próbek do rejestracji danych) istotna rola doktorantki wiązała się z osobistym prowadzeniem badań (pod opieką wykwalifikowanego mikrobiologa) oraz z udoskonalaniem i modyfikacją układu optycznego. Na etapie merytorycznej analizy danych (selekcja/redukcja cech, budowa klasyfikatorów, badania eksperymentalne) wykorzystywane było oprogramowanie do analizy obrazów ImageJ oraz środowisko R (z autorskimi makrami).

Rozdziały 4 (Dyskusja i wyniki) i 5 (Podsumowanie) zawierają syntetyczne przedstawienie zrealizowanych zadań badawczych i uzyskanych rezultatów ze szczególnym uwzględnieniem zalet opracowanej metody rozpoznawania bakterii.

Przedstawiona rozprawa prezentuje najważniejsze rezultaty objęte jej tematyką w logicznym układzie oraz w szerokim kontekście aktualnego stanu wiedzy w zakresie metod i technik rozpoznawania gatunku bakterii. Na szczególnie pozytywne podkreślenie zasługuje zestaw znaczących publikacji, zawierających rezultaty prowadzonych badań oraz włączenie ich w formie załączników w strukturę rozprawy. Umożliwiło to zaprezentowanie uzyskanych wyników typowych dla prac koncepcyjno-eksperymentalnych w sposób szczegółowy i kompletny. Ogólnie zatem, objętość pracy, jej zakres, sposób ujęcia materiału oraz redakcję całości oceniam pozytywnie.

3. Uzyskane wyniki

Celem pracy było wykazanie następujących tez (cyt.):

1. *Uwzględnienie w procesie identyfikacji naturalnego podziału widm dyfrakcyjnych na obszary oraz średnicy widm ma istotny wpływ na możliwości identyfikacji bakterii.*

2. Wyznaczenie na podstawie wartości pikseli właściwych cech liczbowych, uwzględniających charakterystyczną teksturę i morfologię zarejestrowanych widm dyfrakcyjnych Fresnela kolonii bakterii zapewnia optymalne wykorzystanie informacji przez nie niesionej.
3. Wybór odpowiednich cech liczbowych spośród cech ekstrahowanych z widm dyfrakcyjnych Fresnela kolonii bakterii pozwala na budowę optymalnych modeli klasyfikacyjnych.
4. Zastosowanie normalizacji zarejestrowanych widm dyfrakcyjnych Fresnela kolonii bakterii zapewnia ich dokładniejsze porównanie, co prowadzi do zmniejszenia wartości błędu identyfikacji.
5. Metoda cechuje się dokładnością, która umożliwia identyfikację na poziomie serowarów bakteryjnych.
6. Wprowadzenie nieznacznych modyfikacji układu optycznego zapewnia powtarzalność pomiarów z zaledwie nieznacznym pogorszeniem otrzymanych wyników.
7. Zastosowanie właściwych sposobów i algorytmów klasyfikacji i identyfikacji oraz weryfikacji otrzymanych wyników pozwala na uzyskanie rezultatu identyfikacji bakterii z najmniejszym możliwym błędem.

Wykazanie tak nakreślonych tez doprowadziło do realizacji szeregu szczegółowych zadań badawczych i do konkretnych osiągnięć, które – w ujęciu etapowym przedstawionym w pracy – są następujące (wyniki każdego z etapów zostały przedstawione w oddzielnej publikacji):

1. I etap prac badawczych polegał na rozwiązaniu i przeanalizowaniu zadania automatycznego rozpoznawania 4 gatunków bakterii (*Salmonella enteritidis*, *Staphylococcus aureus*, *Citrobacter freundii*, *Proteus mirabilis*). W pierwszej kolejności dokonano ekstrakcji liczbowych cech pierwotnych reprezentujących widma dyfrakcyjne rozpoznawanych kolonii bakterii. Zaproponowano tutaj ciekawą koncepcję podziału widma na rozłączne obszary o kształcie współśrodkowych pierścieni i przyjęciu dla każdego pierścienia dwóch cech: średniej i odchylenia standardowego wartości pikseli pierścienia, rejestrowanych w 8 bitowej rozdzielczości. Liczba cech była w eksperymentach zmieniana poprzez zmianę liczby pierścieni (przyjęto 3, 5 i 10 pierścieni). Następnie, dla każdego przypadku wyznaczone cechy pierwotne poddano redukcji z zastosowaniem przekształcenia PCA. Wyniki przekształcenia nie zostały dalej wykorzystane w zadaniu klasyfikacji, ale posłużyły jedynie do wizualizacji materiału empirycznego (dla dwóch składowych głównych). W zadaniu rozpoznawania zastosowano klasyfikator liniowy z pierwotnym wektorem cech dla różnej liczby pierścieni. Przedstawione macierze konfuzji pozwalają na przeanalizowanie popełnianych błędów. Oceniona została jakość klasyfikacji (dla 6, 10 i 20 cech) z wykorzystaniem metody walidacji krzyżowej (CV) (w domyśle – jednego przebiegu metody) – najlepszy rezultat uzyskano dla 20 cech (7.96% błędnych rozpoznań). W drugiej części badań przeprowadzono selekcję cech z wykorzystaniem metody ANOVA stanowiącej metodę z grupy tzw. filtrów. Metody te pozwalają na ocenę poszczególnych cech i sporządzenie ich rankingu pod względem określonego kryterium oceniającego „siłę dyskryminującą” cechy. Następnie powtórzono rozpoznawanie 4 gatunków bakterii z zastosowaniem klasyfikatora liniowego, ale wykorzystującym cechy według sporządzonego rankingu. W publikacji nie został przytoczony pełen rezultat obejmujący jakość klasyfikacji dla kolejno dokładanych cech. Określono jedynie jakość klasyfikacji

- dla najlepszego zestawu cech, dla każdego z 3 badanych przypadków – najlepszy rezultat uzyskano dla 10 pierścieni i 14 cech (5.97% błędnych rozpoznań).
2. Na II etapie badań rozszerzono analizowane zadanie rozpoznawania bakterii o 3 nowe gatunki (*Escherichia coli*, *Staphylococcus intermedius*, *Pseudomonas aeruginosa*), czyli do problemu 7-mio klasowego. Cechy opisujące widmo pozyskano według powtórzonej koncepcji podziału obrazu na współśrodkowe pierścienie przyjmując – jak poprzednio – dla każdego pierścienia jako cechy średnią i odchylenie standardowe wartości pikseli. Podobnie jak poprzednio, zastosowano przekształcenie PCA do wizualizacji danych, a do rozpoznawania wykorzystano klasyfikator liniowy z cechami wyznaczonymi według rankingu wynikającego z metody ANOVA. Najlepszy rezultat uzyskano dla 10 pierścieni i 15 cech (8.11% – wynik gorszy niż poprzednio, co wynika z większej liczby klas).
 3. III etap badań obejmował – jak poprzednio – analizę zadania rozpoznawania 7 gatunków bakterii, ale ze zdecydowanie bogatszym warształem metodologicznym pozwalającym na uzyskanie nowych, interesujących z praktycznego punktu widzenia, wyników. Nowe aspekty badań są następujące:
 - Wprowadzenie normalizacji obrazów, mającej na celu ujednoczenie skali szarości obrazów widm dyfrakcyjnych kolonii bakterii. Jak wynika z badań eksperymentalnych, uwzględnienie normalizacji prowadzi do wyraźnej poprawy jakości klasyfikacji dla wszystkich analizowanych źródłowych wektorów cech.
 - Zastosowanie dwóch nowych klasyfikatorów w zadaniu rozpoznawania gatunków kolonii bakterii: klasyfikatora kwadratowego (QDA) z kwadratowymi funkcjami klasyfikującymi oraz klasyfikatora SVM (maszyna wektorów wspierających). Oba klasyfikatory znacząco powiększają potencjał klasyfikacyjny w porównaniu z algorytmem LDA stosowanym na poprzednich etapach badań.
 - Rozszerzenie kryteriów oceny badanych metod klasyfikacji o dwa wskaźniki: czułość i specyficzność. W przypadku wieloklasowym, jaki występuje w analizowanych zadaniach rozpoznawania, oba wskaźniki są wyznaczone jako wartości uśrednione zestawu klasyfikatorów binarnych (klasa kontra reszta). Nowe wskaźniki pozwalają na kompleksową ocenę badanych metod klasyfikacji. W tym kontekście, najlepszy okazał się klasyfikator QDA z cechami pochodzącymi z podziału widm dyfrakcyjnych na 10 pierścieni i z przeprowadzoną normalizacją. Uzyskany wynik wynosi 3.14% błędnych klasyfikacji przy czułości 100% i specyficzności 96.35%.
 - Przeprowadzenie rozpoznawania z wykorzystaniem metod LDA, QDA i SVM w oparciu o cechy uzyskane z rankingu wyznaczonego w procedurze ANOVA. W tym przypadku, najlepszy rezultat uzyskany został dla algorytmu QDA i 13 cech z normalizacją obrazów widm dyfrakcyjnych i wynosił on 1.43% błędnych rozpoznań przy czułości 98.26% i specyficzności 98.67%
 4. W IV etapie badań analizowano problem rozpoznawania 7 gatunków bakterii dla podziału widm dyfrakcyjnych na 10 rozłącznych pierścieni. Dla każdego pierścienia wyznaczono teraz 7 pierwotnych cech, które obejmowały – jak poprzednio – średnią i odchylenie standardowe wartości pikseli i dodatkowo skośność, kurtozę, względną gładkość rozkładu, jednorodność rozkładu i entropię. Wybór cech dokonany zgodnie z wyznaczonym rankingiem i ich zastosowanie w klasyfikatorze QDA doprowadziło do najlepszego rezultatu równego 0.857% błędnych rozpoznań przy czułości 98.2% i specyficzności 99.33%.

5. Podstawowym wynikiem V etapu jest zestawienie rezultatów uzyskanych przez doktorantkę w dwóch poprzednich etapach z rezultatami zespołu z Purdue University w USA. Bezpośrednia analiza porównawcza jest niemożliwa ze względu na różne sformułowane zadania rozpoznawania i różne zbiory danych. Nie mniej jednak, uzyskanie liczbowo lepszych rezultatów dla bardziej złożonego zadania rozpoznawania (różnicowanie 7 gatunków bakterii w porównaniu z 4 gatunkami różnicowanymi przez zespół konkurencyjny) korzystnie świadczy o opracowanej metodzie rozpoznawania gatunków bakterii.
6. VI etap prac stanowi z jednej strony podsumowanie opracowanej metody automatycznego rozpoznawania bakterii na podstawie widm dyfrakcyjnych, a z drugiej obejmuje optymalizację wszystkich etapów metody (przygotowanie próbek, układ optyczny, analiza danych) pod kątem jej komercjalizacji i stosowania w praktyce mikrobiologicznej. Na bazie nowego zbioru danych zgromadzonego dla problemu 5-cio klasowego (*Escherichia coli*, *Staphylococcus aureus*, *Proteus mirabilis*, *Salmonella enteritidis* oraz *Salmonella typhimurium*) przeprowadzono badania obejmujące testy czasów inkubacji wpływających na średnicę wyhodowanej kolonii, pozwalające dobrać parametry układu optycznego (m.in. odległość badanych kolonii bakterii od obiektywu kamery rejestrującej widma dyfrakcyjne), a także analizę metody losowania warstwowego w celu zwiększenia reprezentatywności prób do walidacji krzyżowej. Do selekcji cech, obok analizy wariancji, zastosowano dywergencję Fishera, jako metody pozwalające ocenić przydatność cechy do różnicowania klas (gatunków bakterii) i uporządkować cechy w formie odpowiedniego rankingu. Do rozpoznawania wykorzystano klasyfikatory QDA i SVM, a do oceny jakości klasyfikacji (dla różnych badanych i optymalizowanych parametrów) 50 powtórzeń 10-cio krotnej walidacji krzyżowej.

Przeprowadzone badania ujęte w 6 etapów pokazują kolejne wzbogacanie stosowanych metod i technik analizy danych (w szczególności metod selekcji cech i algorytmów klasyfikacji) i rozwój warsztatu badawczego. Stanowią one określoną całość dotyczącą eksperymentalnego badania opracowanej metody automatycznego rozpoznawania bakterii na podstawie widm dyfrakcyjnych, ukierunkowanych na doświadczalne wykazanie prawdziwości sformułowanych tez badawczych.

4. Uwagi szczegółowe

W recenzowanej rozprawie dostrzega się usterki, które można ująć w następujące punkty:

1. We wszystkich przeprowadzonych eksperymentalnych analizach porównawczych (klasyfikatorów, metod selekcji cech, itp.) uzyskane wyniki winny być poparte analizą statystyczną, co jest standardem w tego typu badaniach. Dopiero potwierdzenie, iż między wynikami jakości dwóch porównywanych algorytmów istnieje statystycznie istotna różnica (na przyjętym poziomie istotności) pozwala na wyciągnięcie wniosku o wyższości jednego algorytmu na drugim. Jeśli dodatkowo porównywane są wyniki oddzielnie dla różnych zbiorów danych, można skorzystać z odpowiednich testów rankingowych (przykłady takich testów można znaleźć w J. Demsar, Statistical comparison of classifiers over multiple data sets, *Journal of Machine Learning Research* 7 (2006) 1–30).
2. Zastosowanie w rozpoznawaniu kolonii bakterii klasyfikatorów typu LDA, QDA oraz SVM stanowi niewątpliwie wartościowy zakres badań, ale trudno na tej podstawie uznać, że uzyskano najlepszy rezultat (teza nr 7), lub że dokonano budowy optymalnych modeli

klasyfikacyjnych (teza nr 3). Metody i techniki klasyfikacji i wynikające z nich algorytmy są bardzo bogate. Niektóre sugestie dotyczące zastosowania innych podejść mogą wynikać z uzyskanych i przedstawionych wyników. Np. wizualizacja danych na podstawie PCA przedstawiona w pracy (6) sugeruje, że skuteczne powinno być tutaj binarne drzewo decyzyjne, zbudowane na bazie dwóch składowych głównych. Z kolei czułość wyników na rodzaj zastosowanego klasyfikatora pozwala przypuszczać, iż warto tu zastosować systemy wieloklasyfikatorowe z oceną kompetencji klasyfikatorów bazowych.

3. Powyższą uwagę można odnieść do zastosowanej w pracy bardzo konsekwentnie metody ANOVA selekcji cech. Jak wynika z b. bogatej literatury dotyczącej metod selekcji cech, warto zainteresować się metodami typu *wrapper*, które mają tę własność (będącą zarówno wadą jak i zaletą), że są dedykowane konkretnemu klasyfikatorowi. Innymi słowy, dokonany wybór (selekcja) cech jest oceniana jakością działania klasyfikatora. Recenzent sugeruje, aby przeprowadzić stosowne badania eksperymentalne przy ewentualnej kontynuacji prac badawczych.
4. Niezręczne bądź niewłaściwe sformułowania: *średnia i odchylenie standardowe (zerowy i pierwszy centralny moment statystyczny)* (str. 33), *najbardziej optymalny* (str.36), dyskusyjna definicja klasyfikacji na str. 19.

5. Podsumowanie recenzji – ocena rozprawy

Reasumując stwierdzam, iż Pani mgr inż. Agnieszka Suchwałko wykazała się dużą wiedzą z zakresu zaawansowanego przetwarzania danych pomiarowych, obejmującego w szczególności analizy statystyczne, metody zmniejszania wymiarowości danych (redukcja i selekcja), algorytmy klasyfikacji oraz techniki eksperymentalnej oceny jakości algorytmów, a także opanowaniem i właściwym posługiwaniem się doświadczalnym warsztatem eksperymentowania w obszarze mikrobiologii. Przedstawiona praca zawiera poprawnie sformułowany i rozwiązany problem badawczy i stanowi istotny wkład w dziedzinę inżynierii biomedycznej w obszarze metod i technik automatycznego rozpoznawania gatunków bakterii. Zawarte w niej rezultaty obejmujące akwizycję obrazów widm dyfrakcyjnych kolonii bakterii z wykorzystaniem dedykowanego układu optycznego, opracowanie metod parametryzacji widm, zastosowanie różnych technik redukcji i selekcji cech oraz klasyfikatorów, a także przeprowadzenie eksperymentalnych analiz porównawczych są oryginalne i zostały przedstawione w sposób wyczerpujący na odpowiednim poziomie matematycznych formalizmów. Uważam, że praca doktorska Pani mgr inż. Agnieszki Suchwałko stanowi oryginalne rozwiązanie problemu naukowego, a tym samym spełnia wymogi Art.13 pkt. 1 Ustawy o stopniach naukowych i tytule naukowym z dnia 14.03.2003 r. Wniosuję o dopuszczenie Pani mgr A. Suchwałko do publicznej obrony.

