Prof. Leszek Trybus
Katedra Informatyki i Automatyki
Politechnika Rzeszowska

Rzeszów, 27.01.2017 r.

Recenzja
rozprawy doktorskiej mgr inż. Marcina Ochaba
pt. *System wspomagania decyzji dla wczesnego wykrywania dysplazji oskrzelowo-płucnej u wcześniaków*

1. Zakres i cele

Dysplazja oskrzelowo-płucna BPD (*bronchopulmonary dysplazia*) jest przewlekłą chorobą płuc wcześniaków o bardzo małej masie ciała (500...1500 g). Charakteryzuje ją wysoka śmiertelność sięgająca 30%. Dysplazję BPD diagnozuje się standardowo po 4. tygodniach od urodzenia na podstawie zespołu parametrów (cech), takich jak wiek płodowy, saturacja hemoglobiny tlenem we krwi, użycie surfakantu, przetrwały przewód tętniczy i innych. Dopiero po tej diagnozie przystępuje się do leczenia.

Nie na jednak przeszkód, by wspomniane parametry wykorzystać do wcześniejszego prognozowania BPD. Medycyna wykorzystuje do tego statystyczną metodę regresji logistycznej RL, która badając wzajemne związki między parametrami pozwala przewidzieć wartość jednego z nich na podstawie wartości pozostałych. Wynikiem jest prawdopodobieństwo trafnej diagnozy. Według specjalistów, miarodajną predykcję przyszłej diagnozy można byłoby sformułować już w tydzień po urodzeniu. Na razie brak jednak specjalistycznego oprogramowania, które na podstawie reprezentatywnej bazy wiedzy gromadzącej dane o wcześniakach z BPD byłoby w stanie formułować takie prognozy.

W środowiskach akademickich zajmujących się klasyfikacją za jedną z najwartościowszych uważa się metodę wektorów wspierających SVM (*support vector machine*). Nie próbowano jej jednak jeszcze stosować do prognozowania dysplazji BPD. Interesująca byłaby więc próba porównania regresji RL z metodą SVM i zbadanie, w jakiej sytuacji korzystniejsza jest jedna albo druga. Powinno to być jednak oparte na solidnie udokumentowanej bazie wiedzy i brać pod uwagę zróżnicowaną liczbę parametrów, które w odniesieniu do danego wcześniaka mogą być dostępne.

Dlatego celem, który postawił przed sobą mgr M. Ochab było opracowanie systemu informatycznego przeznaczonego do wspomagania decyzji dla wczesnego wykrywania dyplacji oskrzelowo-płucnej u wcześniaków. Przyjęto, że prognoza dysplazji BPD będzie formułowana w tydzień po urodzeniu. Na podstawie wykazu dostępnych parametrów system powinien podawać trafność prognozy (w procentach), określonej zarówno metodą RL jak i SVM, odchylenie standardowe oraz dwa inne istotne wyniki (czułość, specyficzność). Cel ten uważam za jak najbardziej zasadny, zarówno pod względem badawczym, jak i utylitarnym.

2. Przegląd treści

We *Wstępie*, na bazie szczegółowego przeglądu literatury medycznej, Autor scharakteryzował dysplazję oskrzelowo-płucną, czynniki ryzyka oraz modele predykcyjne BPD w zależności od wieku płodowego, saturacji hemoglobiny itd. (j.w.). Sformułował również cel rozprawy i tezę.

Rozdział 1, *Dane*, przedstawia źródła danych, podział danych na statyczne (jednorazowe) i dynamiczne (ciągłe, np. tętno) oraz analizę statystyczną parametrów (cech). Dane pochodzą z Oddziału Intensywnej Terapii Noworodka Collegium Medicum UJ, z lat

2004-2010 i dotyczą 109. wcześniaków. Podano histogramy parametrów, zbadano ich współliniowość oraz korelację.

Metody predykcji zastosowane w pracy, tzn. regresję logistyczną RL i metodę wektorów wspierających SVM, a także porównanie sposobów ich działania opisano w rozdz. 2. W odniesieniu do SVM zasygnalizowano problem dualny (rozwiązywany), podano funkcję decyzyjną oraz funkcje podobieństwa wektorów wspierających (sigmoidę i RBF). Przedyskutowano także rolę parametrów. Ponieważ wynikiem działania regresji RL jest prawdopodobieństwo (dla odpowiedniego progu), a w metodzie SVM dwuwartościowa klasyfikacja, więc porównania metod dokonuje się na gruncie trafności, czułości i specyficzności (*sensitivity*).

W rozdz. 3 przedstawiono *Wstępne badania* klasyfikacji przeprowadzone w środowisku Matlab. Przy pomocy eliminacji Jackknife oraz walidacji krzyżowej wykluczono sytuację, w której klasyfikator mógłby „nauczyć się na pamięć" określonych przypadków. Zbadano następnie klasyfikatory RL, które dla liczby parametrów nie przekraczającej 5 zachowywały się korzystniej, z trafnością predykcji rzędu 82%. Metoda SVM zaimplementowana w Matlabie jako funkcja *svmtrain* dawała gorsze wyniki (80%). Sięgnięto więc po bibliotekę LIBSVM (Chang, Lin), co wyraźnie poprawiło wyniki. Stwierdzono, że przy liczbie parametrów do 5, regresja RL daje trafniejszą prognozę, natomiast począwszy od 6 przeważa SVM.

System wspomagania decyzji stanowiący główne osiągnięcie rozprawy jest opisany w rozdz. 4. Dopuszcza on do 14 parametrów charakteryzujących wcześniaki. Przedstawiono budowę bazy wiedzy powstałej w wyniku czasochłonnych obliczeń dla 2^{14} kombinacji parametrów. Podano tabelę z 40. klasyfikatorami, najlepszymi pod względem trafności predykcji. Scharakteryzowano również budowę systemu będącego aplikacją Webową, napisaną w języku PHP z bazą danych MySQL. Podano przykład odpowiedzi systemu dla 6. parametrów (SVM nieco trafniejszy).

Eksperymenty ilustrujące działanie stworzonego systemu są opisane w rozdz. 5. Chodzi w nich o sekwencyjną rozbudowę modeli rozpoczynając od jednoparametrowego, ale przy różnych parametrach startowych. Za pomocą funkcji *sequential fs* Matlaba ustalono, które parametry spośród założonych 14. są najistotniejsze dla trafności predykcji. Odniesiono się również do mniej istotnych.

Wnioski z rozprawy są zawarte w rozdz. 6. Należy do nich wskazanie, które parametry są istotne dla modeli 2-, 3- i 4- parametrowych oraz podkreślenie, że duże znaczenie dla poprawy trafności predykcji miałoby uwzględnienie nastawy respiratora F_iO_2. Wspomniano także o możliwościach rozbudowy systemu.

W *Dodatku* podano przykład liczbowy predykcji dysplazji BPD metodami RL i SVM.

3. Ocena ogólna

Do klasyfikacji danych medycznych dotyczących dysplazji BPD mgr M. Ochab wykorzystał regresję logistyczną RL oraz metodę wektorów wspierających SVM. Pierwsza z nich była już stosowana wielokrotnie, więc należy ją uznać za klasyczną, natomiast zastosowanie SVM można uważać za nowatorskie. Alternatywnie, nadawałyby się również do tego klasyfikatory wykorzystujące sieci neuronowe (o czym wspomniano), drzewa decyzyjne, czy sieci Bayesa.

Omawiając w p. 2.2 metodę SVM Autor scharakteryzował ideę polegającą na maksymalizacji marginesu wyznaczonego przez hiperpłaszczyznę separującą, rozdzielającą dane należące do dwóch klas ze zbioru {-1,1}. Poprawnie sformułował problem pierwotny, którego zwykle się nie rozwiązuje oraz problem dualny, który rozwiązany jest numerycznie jako zadanie programowania kwadratowego QP z ograniczeniami, stosując metodę mnożników Lagrange'a. W skrócie przedstawił podstawowe idee związane z poszukiwaniem rozwiązania dla problemu z danymi liniowo separowalnymi, a następnie jego uogólnienie dla

danych liniowo nieseparowalnych, czyli dla przypadku najczęściej spotykanego w praktyce, także rozpatrywanego w rozprawie. Autor nie przedstawił wyprowadzenia problemu dualnego, zapewne dlatego, że można je znaleźć w wielu pracach, począwszy od twórców algorytmu SVM (Vapnik, Vapnik, Cortes). Wystarczająco dobrze scharakteryzował ważny pomysł polegający na zastosowaniu dodatnio półokreślonej funkcji jądra (tzw. *kernel trick*), dzięki czemu udaje się precyzyjnie sformułować problem optymalizacji dla dowolnych danych liniowo nieseparowalnych. Wyjaśniono zatem, że w praktyce unika się przez to konieczności znajdywania nieliniowego przekształcenia, które transformuje oryginalne dane liniowo nieseparowalne do przestrzeni cech o znacznym wymiarze, gdzie istnieje duża szansa, że obrazy tych danych będą liniowo separowalne. Autor wyjaśnił też ideę wprowadzenia zmiennych swobodnych (tzw. *slack variables*). Nie wdając się w wyprowadzenia, omówił także najważniejsze algorytmy klasyfikacyjne z rodziny SVM, tzn. C-SVC oraz nu-SVC.

Oceniając generalnie, mgr M. Ochab wykazał się wystarczająco dobrą znajomością metody SVM, jak również oprogramowania implementującego ją w środowisku Matlab, a także efektywniejszej biblioteki LIBSVM. Na podkreślenie zasługuje rzetelne przygotowanie bazy danych obejmujące analizy statystyczne, walidację krzyżową itp. System wspomagania decyzji, który opracował może być pomocny we wczesnym wykrywaniu dysplazji BPD.

W spisie literatury znajduje się 5 anglojęzycznych prac, których jest współautorem.

4. Uwagi krytyczne

W odniesieniu do metody SVM pewien niedosyt budzi brak nieco szerszego omówienia metody minimalnej optymalizacji sekwencyjnej SMO wprowadzonej przez Platta, a stanowiącej receptę dla dużych zbiorów danych, gdy wymiar hasjanu dla dualnego problemu SVM staje się sporym ograniczeniem. Jednak usprawiedliwieniem dla Autora może być fakt, że rozpatrywane w rozprawie dane nie są duże, obejmują bowiem tylko 109 rekordów opisujących wcześniaki. Ponadto wybierając SVM do klasyfikacji danych należy wybrać zarówno rodzaj funkcji jądra (funkcja Gaussa, sigmoida, RBF, wielomian itp.), jak również jej parametry (gamma, jak w doktoracie). Dodatkowo należy wybrać optymalną wartość stałej C ograniczającej od góry współczynniki Lagrange'a. W algorytmie przedstawionym na s. 61 optymalizuje się tylko szerokość funkcji Gaussa, bez optymalizacji stałej C, co należałoby wyjaśnić.

Narzędzia wykorzystane do budowy systemu, tzn. język PHP i baza MySQL, choć odpowiednie, są już dość tradycyjne. W pracy brak krótkiego uzasadnienia, dlaczego dokonano takiego wyboru.

Pod względem edytorskim rozprawa jest przygotowana bez zarzutu. Usterek redakcyjnych praktycznie nie ma (wyjąwszy „hemoglogina" na str. 10 (kilkakrotnie) i „wartoś" na s. 54).

5. Wniosek końcowy

Sądzę, że cel postawiony w rozprawie został osiągnięty. Autor opracował system wspomagania decyzji dla wczesnego wykrywania dysplazji BPD formułując prognozę w tydzień po urodzeniu. Oprócz ugruntowanej w środowisku medycznym metody regresji logistycznej, zastosowano metodę wektorów wspierających SVM dającą trafniejsze prognozy przy większej liczbie parametrów. Wykorzystanie metody SVM do klasyfikacji dysplazji BPD należy uznać za nowatorskie.

Uważam, że rozprawa doktorska mgr inż. Marcina Ochaba pt. *System wspomagania decyzji dla wczesnego wykrywania dysplazji oskrzelowo-płucnej u wcześniaków* spełnia warunki określone w Ustawie z 2003 r. o stopniach naukowych i tytule naukowym, i wnoszę o dopuszczenie jej do publicznej obrony.