Predykcja zachorowalności na raka piersi w Polsce

Konrad Czechowski, Kamil Koziej, Bartosz Piotrowski, Jakub Tyrek

Dane

Model predykcyjny zbudowaliśmy w oparciu o następujące dane: wskaźniki zachorowalności w latach 2010 i 2011 oraz wskaźniki urbanizacji w powiatach i grupach wiekowych. Przesłanką uzasadniającą dołączenie wskaźnika urbanizacji do danych było jego istotne skorelowanie z zachorowalnością.

Modele predykcyjne

Naszym celem była predykcja na podstawie danych zachorowalności w grupach wiekowych i powiatach w roku 2012. Przetestowaliśmy kilka popularnych modeli, m. in.:

- model liniowy,
- glmnet,
- lasy losowe,
- kNN,
- SVN.

Żaden z powyższych modeli nie dawał predykcji lepszej, niż proste branie średniej zachorowalności w danej grupie z dwóch lat poprzednich. Porównanie modeli pod względem błędu średniokwadratowego pokazuje wykres słupkowy w prawym górnym rogu. *naive* oznacza model biorący średnią z lat poprzednich, natomiast *modified.knn* oznacza stworzony przez nas model, który opisujemy poniżej. Walidację modeli przeprowadziliśmy techniką 10-krotnej kroswalidacji.

Nasz model

Z powodu niezadowalających wyników dla typowych modeli, skonstruowaliśmy nasz własny, lepiej przystosowany do danych. Jego idea opiera się na następującej obserwacji: w dużych powiatach wariancja zachorowalności w kolejnych latach jest niewielka, więc zasadne jest branie średniej zachorowalności z lat ubiegłych jako predykcję dla roku kolejnego; natomiast w małych powiatach wariancja zachorowalności jest duża (w porównaniu do wielkości populacji), dlatego w ich przypadku taka predykcja obarczona będzie dużym błędem.

Model działa następująco:

Niech G będzie danym podzbiorem populacji z zadaną płcią, przedziałem wiekowym i powiatem, dla której chcemy wyznaczyć predykcję zachorowalności.

1) Jeżeli liczebność grupy G przekracza zadany próg k_W, model zwraca jako predykcję średnią zachorowalność lat poprzednich w grupie G.
2) W przeciwnym razie do danej grupy dobieramy N grup z tego samego województwa i przedziału wiekowego, które mają jak najbliższe wskaźniki urbanizacji, aż do momentu, gdy suma populacji w tych grupach przekroczy zadany próg k_W. Jako predykcję zwracamy wówczas średnią ważoną zachorowalności w tych grupach w latach poprzednich.
3) Próg k_W jest zoptymalizowany dla każdego W = <płeć, przedział wiekowy> osobno. Przykłady analizy optymalnego k_W dla konkretnych grup obrazują dwa wykresy po prawej na środku.