

Streszczenie

Szpiczak mnogi jest nowotworem złośliwym szpiku kostnego, którego rozwój następuje w wyniku mutacji komórek plazmatycznych o złożonej patogenezie. Jest on nadal chorobą nieuleczalną, a głównym celem leczenia jest wydłużenie czasu przeżycia. Ważnym więc zadaniem staje się jego wczesna diagnoza, co jest szczególnie istotne przy wykryciu zachorowań w II i III stadium. Szpiczak mnogi jest chorobą trudnodiagnozowalną, dlatego pojawiają się próby wykorzystania rozwoju technologii obliczeniowych w postaci narzędzi sztucznej inteligencji do przetwarzania i analizy danych dla opracowania algorytmów, pozwalających na wybór istotnych oznak świadczących o wystąpieniu tej choroby. Główną metodą wykorzystywaną do tego celu będzie uczenie maszynowe.

W rozprawie przedstawione zostały metody analizy niepełnych danych wraz z metodami pozwalającymi na ich odzyskanie. Jest to etap wstępny analizy danych, który jest niezbędny dla prawidłowo realizowanych badań. Wstępne przetwarzanie danych ma kluczowe znaczenie dla prawidłowej analizy i interpretacji wyników, zwłaszcza podczas badania złożonych chorób, takich jak szpiczak mnogi. Ponieważ dane cechuje wysoka wymiarowość, dlatego w pracy omówione zostały metody selekcji cech stosowanych w uczeniu maszynowym. Dla oceny relacji między cechami a zmienną docelową, wykorzystuje się metryki takie jak entropia, entropia warunkowa, informacja wzajemna i dywergencja Kulbacka-Leiblera. Analiza wykazała dużą zależność tych metod od charakterystyki informacyjnej analizowanych danych.

W rozprawie przeprowadzono badania zbioru danych eksperymentalnych, zabranych wśród pacjentów szpitala klinicznego w Lublinie. Zrealizowano wstępne przetwarzanie danych, obejmujące uzupełnienie brakujących wyników. Na kolejnym etapie przetwarzania danych, przeprowadzona została selekcja cech, w wyniku której z zestawu 283 cech uzyskano redukcję do 20 cech. Do oceny jakości wybranych podzbiorów wykorzystano dwanaście klasyfikatorów.

W sumie przeprowadzone zostały 1534 eksperymenty obliczeniowe, testujące różne kombinacje, 9 metod selekcji (między metodami filtracji, zwijania i osadzania) oraz 12 różnych klasyfikatorów. Jakość każdego modelu została oceniona przy pomocy takich miar jak: dokładność, czułość, swoistość, F-score i parametry AUC (area under the curve).

Na podstawie przeprowadzonych obliczeń do klasyfikacji (diagnostyki) cech szpiczaka mnogiego wybrano dwa modele JMI-B-20 i NJMIM-NB-20, gdyż uzyskały one najwyższą dokładność i działają z mniejszą ilością cech. W końcowej części pracy przedstawiono propozycję dalszych badań dotyczących diagnostyki szpiczaka mnogiego.

Słowa kluczowe: szpiczak mnogi, przetwarzanie danych, metody i algorytmy klasyfikacji, cechy diagnostyczne