22.01.2019

Błażej Kucman

ind. 238228

gr.1

Inteligencja Obliczeniowa – Projekt 4

**Przetwarzanie obrazu**

Tematem zadania było dokonanie klasyfikacji obrazów za pomocą wybranych klasyfikatorów.

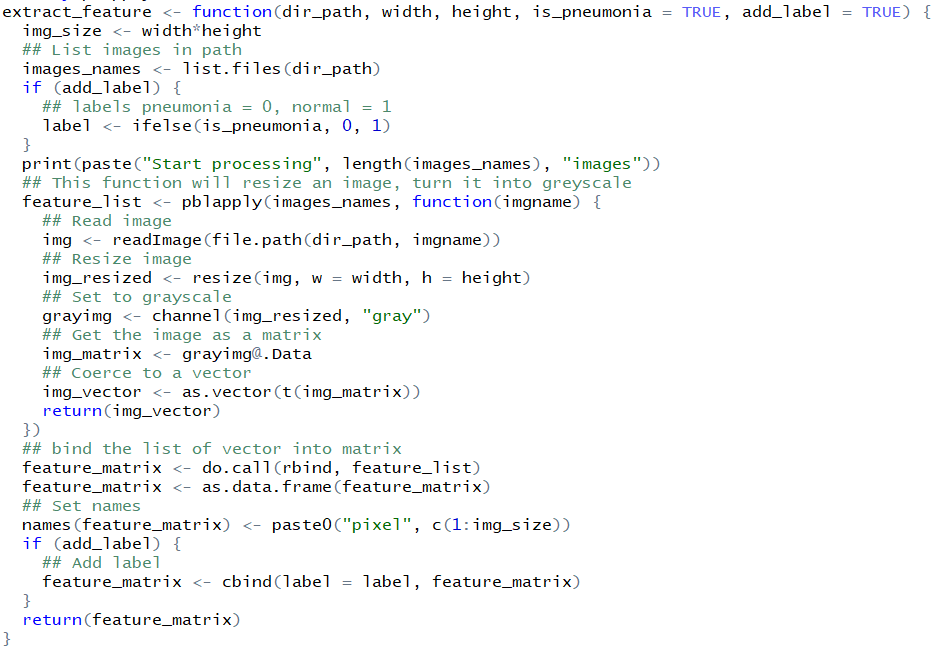
Wybrane klasyfikatory: CNN, Random Forest, Gradient boosting i SVM.

Baza obrazów

Wybrana została baza zawierająca zdjęcia RTG płuc z których część przedstawiała zapalenie płuc a część nie. Zdjęcia w bazie były w formacie jpeg, co powoduje konieczność przetworzenia ich do postaci wektorowej. Wektor ten zawiera kolumnę label (zdrowy, chory) i 784 wartości, z których każda reprezentuje pojedynczy pixel. Zdjęcia przed przeliczeniem zostały przeskalowane do rozmiaru 28x28 i odcieni szarości.

Label: 0 – zapalenie płuc, 1 - zdrowy

Do konwersji zdjęć została użyta biblioteka EBImage poniżej kod odpowiedzialny za to.

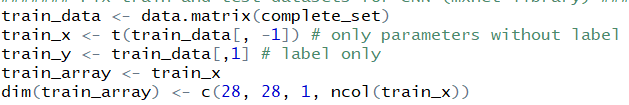


Klasyfikatory

CNN - Convolutional Neural Network

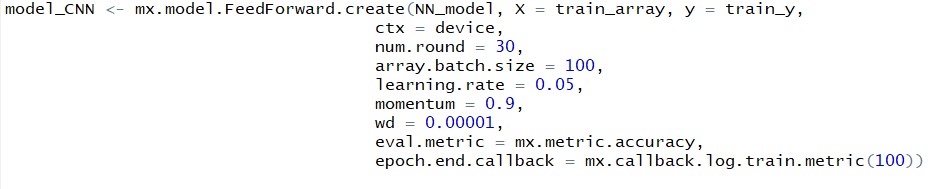
Jest specyficznym typem sieci neuronowej, która wykorzystuje perceptory. Ta splotowa sieć neuronowa ma warstwę wejściową i wyjściową i różne ukryte warstwy czyli wagi wyliczane pod czas uczenia, co znaczy że są to współczynniki, które sugerują ważniejsze parametry po przez swoją wielkość.

Została użyta biblioteka mxnet. Na potrzeby tego klasyfikatora należało przekształcić macierz do postaci , w której zdjęcie jest przedstawione jako pod obiekt macierzy o rozmiarze 28x28. Oraz oddzielić parametry od label.



Sieć do uczenia została przygotowana opierając się o przykładowym toturialu.

Tworzenie modelu



Działa na dwóch warstwach sieci neuronowej. Obliczenia na CPU. Wykonywane jest 30 rund uczenia tendencja rozpoznawania jest wzrostowa wraz z każdą następną. 30 jest wystarczające gdyż nie osiąga już lepszej skuteczności.

|  |  |  |  |
| --- | --- | --- | --- |
| Real | Predicted | | |
|  | 0 | 1 |
| 0 | 385 | 5 |
| 1 | 121 | 113 |

Label: 0 – zapalenie płuc, 1 - zdrowy

Dokładność: 0.7980769

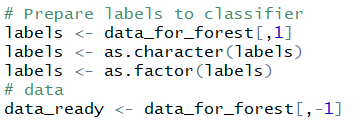
Widzimy tutaj bardzo dobrą rozpoznawalność osób chorych. Ale jednocześnie dużo osób zdrowych   
oznaczonych jako chore. Jak się później okaże ten klasyfikator osiągnął najmniejszą liczbę pomyłek.

Przeprowadzając analizę dokładność względem ilości rund trenowania, oddzielnie od powyższych   
wyników ciężko się dopatrzeć jakiejś zależności. Co można dostrzec to to, że najlepszy wynik został  
otrzymany przy 10 rundach. Powyżej 25 rund można zauważyć tendencję pogarszania się wyniku.

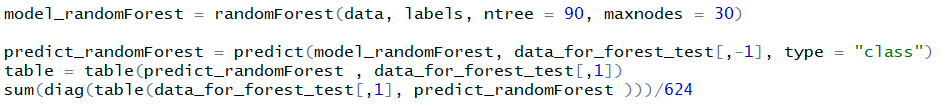
Random Forest

Metoda klasyfikacji polegająca na tworzeniu wielu drzew decyzyjnych na podstawie losowego zestawu danych. Zaletami jest pracowanie nawet na tysiącach cech i działa przy niekompletnych danych.

Dane zostały podzielone na label i parametry. Label wymagało konwersji do factor, w innym wypadku output był by liczbami zmiennoprzecinkowymi.



Użycie klasyfikatora



|  |  |  |  |
| --- | --- | --- | --- |
| Real | Predicted | | |
|  | 0 | 1 |
| 0 | 385 | 5 |
| 1 | 143 | 91 |

Label: 0 – zapalenie płuc, 1 - zdrowy

Dokładność: 0.7628205

Trafione choroby na bardzo dobrym poziomie, ale wynik pomylonych zdrowych z chorymi bardzo zły.

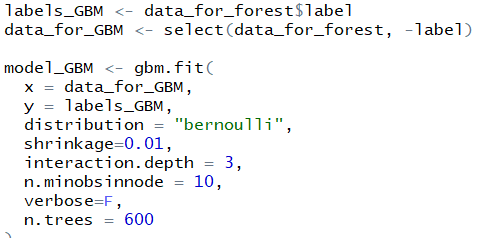
Wynik uzyskany przy ntrees = 30, 4 razy większa liczba drzew dawała wynik lepszy tylko o 1.5 %. Wiec przy tym pomiarze pozostałem.

Ciekawą obserwacją było że przy ntrees = 300 wynik był ~78% a przy 1000 był ~77%. Wykres na którym można zaobserwować ten trend poniżej.

Gradient boosting

Kolejny klasyfikator korzystający z drzew. Na początku drzewa mają równe wagi. Po dokonaniu pierwszy ocen zwiększa wagi tych obserwacji które są trudne do sklasyfikowania i obniża wagi tych, które są łatwe. Następne drzewo powstaje na tych ważonych danych. Obliczany jest błąd takiego drzewo. Proces ten jest powtarzany określoną liczbę razy.

Przygotowanie danych oraz wyznaczenie modelu



Wybrany rozkład to bernoulli zalecany dla outputu 0 i 1. Krok 0.01 .

Predict w przypadku gradient boosting różni się tym, że należy podać liczbę drzew z ilu ma korzystać. Ponieważ może się okazać, wystarczy mniej niż podaliśmy.



|  |  |  |  |
| --- | --- | --- | --- |
| Real | Predicted | | |
|  | 0 | 1 |
| 0 | 387 | 3 |
| 1 | 172 | 62 |

Label: 0 – zapalenie płuc, 1 - zdrowy

Dokładność: 0.7195513

Klasyfikator najlepiej oszacował chorych , ale także wynik pomylonych zdrowych z chorymi jest najgorszy. Mogło by się wydawać że to dobry klasyfikator lecz liczba pomyłek jest nie do przyjęcia patrząc na tabelkę.

Zwiększanie liczby drzew nie wpływało na polepszenie wyniku przy 600 jak i przy 1000 był on praktycznie taki sam.

Dokładność w tym przypadku stabilizuje się przy około 800 drzewach. Lecz już od 600 są to małe różnice. Przy około 1400 jakość na chwilę nawet się nieznacznie pogorszyła.

SVM – Support Vector Machines

Funkcja liniowa separująca wyznacza podział przestrzeni na obszary odpowiadające dwóch klasom decyzyjnym. Działanie polega na mapowaniu danych na wielowymiarową przestrzeń właściwości w sposób umożliwiający kategoryzację punktów danych.

Wyznaczenie modelu



Typem operacji jest wybrana klasyfikacja danych. Problem rozwiązywany jest liniowo. Na zasadzie opisanej wyżej.

|  |  |  |  |
| --- | --- | --- | --- |
| Real | Predicted | | |
|  | 0 | 1 |
| 0 | 386 | 4 |
| 1 | 146 | 88 |

Label: 0 – zapalenie płuc, 1 - zdrowy

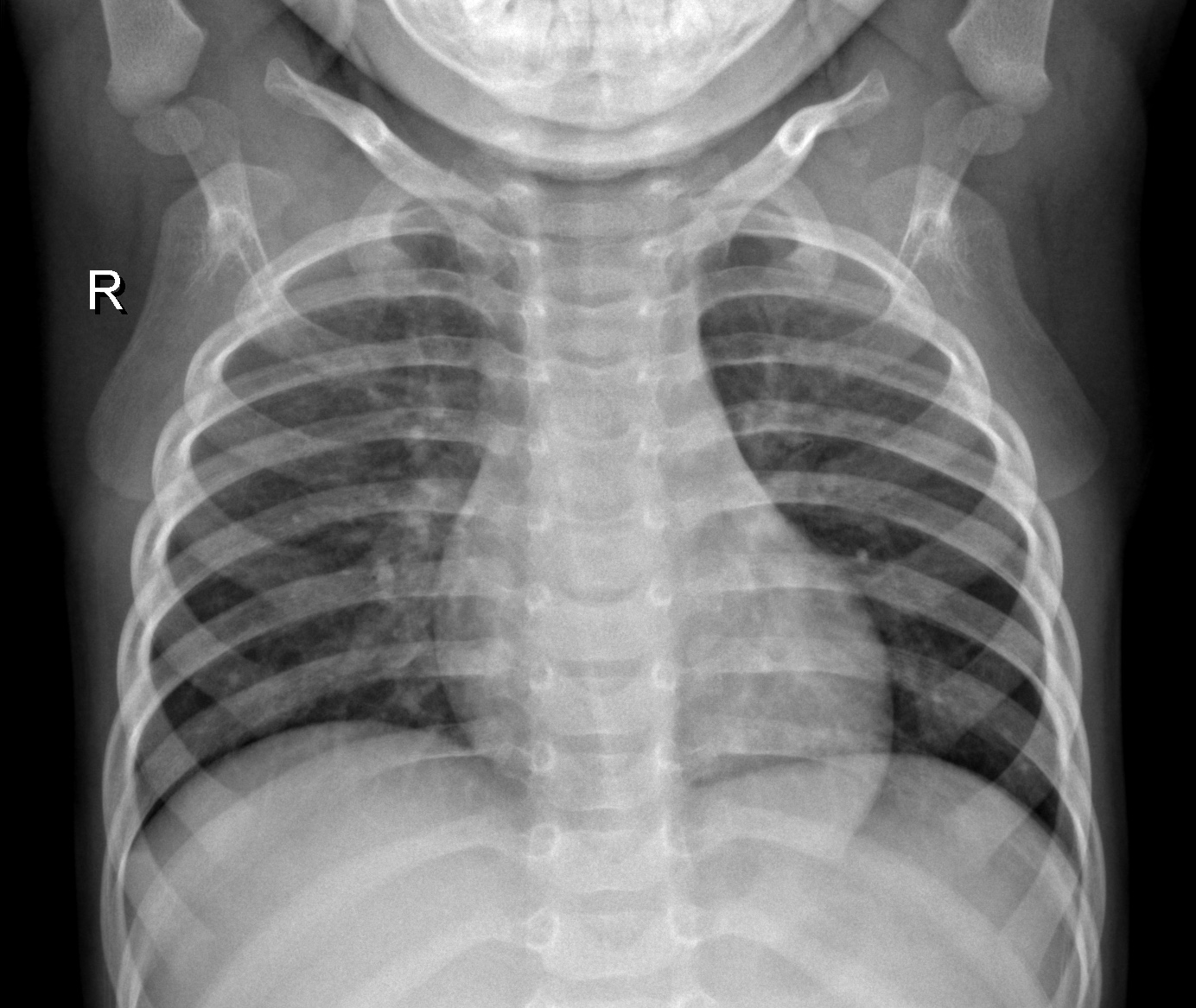
Dokładność: 0.7596154

Trafione choroby na bardzo dobrym poziomie, drugi pod tym względem, ale wynik pomylonych zdrowych z chorymi bardzo zły.

Podsumowanie

Podsumowując wyniki może je przedstawić na wykresie i dobrze widać że najlepiej zadziałał CNN. Wszystkie klasyfikatory wykazały się bardzo dobrym odgadywaniem osób chorych, lecz źle radziły sobie z diagnozowaniem zdrowych, w tym aspekcie najlepszy okazał się CNN a najgorszy GBM.

Jeśli chodzi o rodzaj błędu to większość chorych została by zdiagnozowana co jest zamierzonym efektem. Niestety liczby osób zdrowych sklasyfikowanych jako chore jest nie do przyjęcia gdy nawet dwa razy więcej jest źle sklasyfikowanych niż dobrze w tym aspekcie. Prawdopodobnie może wynikać to z faktu, że baza treningowa zdjęć zawierała 3875 chorych i 1341 zdrowych co spowodowało bardzo dobre rozpoznanie przez klasyfikatory chorych ale nie zdrowych bo było ich po prostu za mało. Kolejnym powodem może być badany problem sam w sobie. Ponieważ część zdjęć chorych jest bardzo podobna do zdrowych poniżej przykład.

Widzimy tu bardzo podbne płuca z podobnymi wcięciami szarości na płucach ale jedne są zdrowe drugie chore. Lew zdjęcie to zapalenie płuc a prawe zdrowe, klasyfiaktor mając więcej danych chorych może te obszary potraktować jako choroba, a może być to po prostu pozostałości po kiedyś przebytej chorobie, poniewaź choroby płuc zostawiają ślady.

Lecz patrząc z innej strony chorych klasyfikuje bardzo dobrze więc przy dodatkowych badaniach można odrzucić tych zdorwych a chorzy będą leczeni. Lepiej tak niż błędnie sklasyfikować chorych.