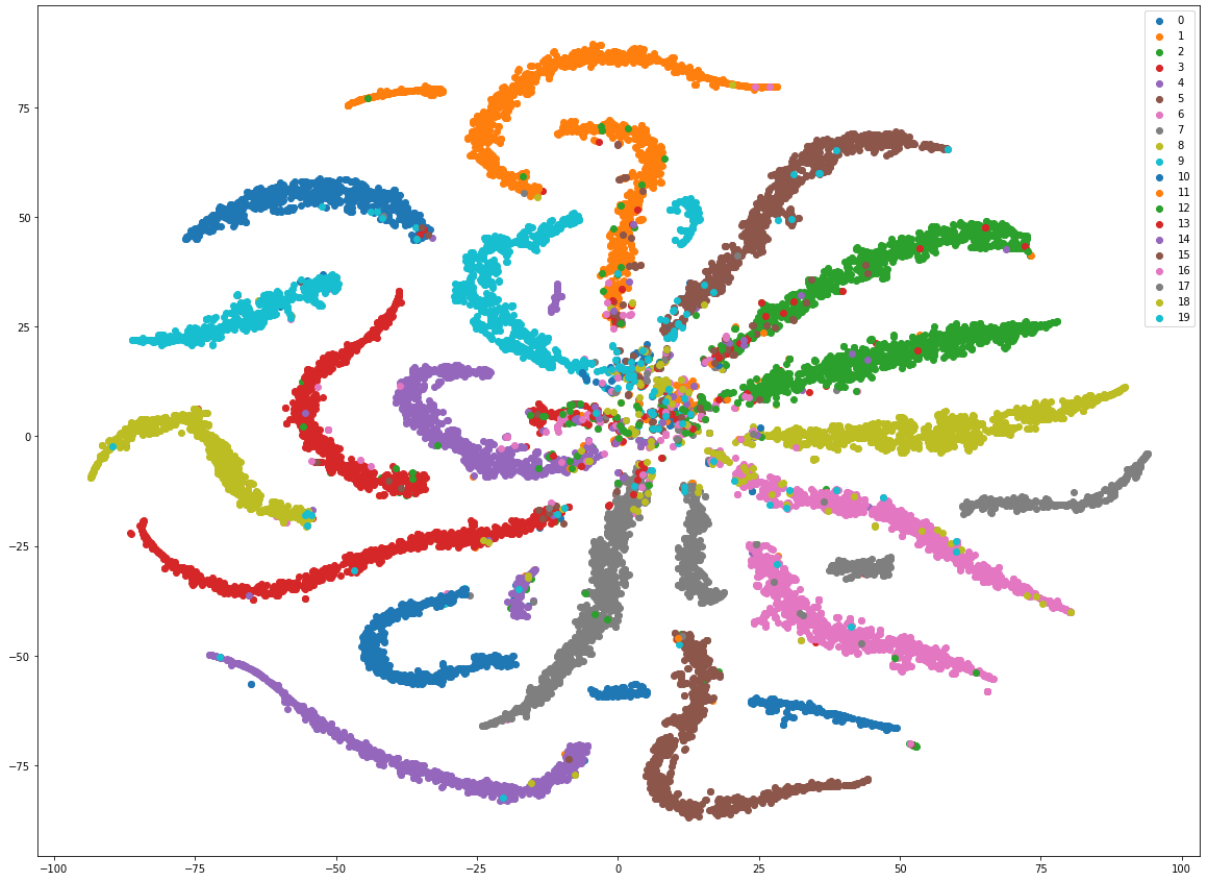
Arkadiusz Kraus
Filip Ślazyk

Raport z realizacji projektu - temat nr 1

1. Eksploracja zbioru danych

W projekcie badaliśmy zbiór danych złożony z 18846 próbek, pochodzących z przetworzenia zbioru TNG. Każda próbka miała 128 cech. W celu lepszego zrozumienia, jaki zbiór badamy, wykonaliśmy wizualizację przy pomocy t-SNE. Zbadaliśmy też liczebność klas.



Rys. 1. Ułożenie poszczególnych klas po zastosowaniu t-SNE

0	799	7	990		
1	973	8	996	14	987
2	985	9	994	15	997
3	982	10	999	16	910
4	963	11	991	17	940
5	988	12	984	18	775
6	975	13	990	19	628

Tab. 1. Liczby próbek w poszczególnych klasach 0-19

Wizualizacja pozwala stwierdzić, iż mamy do czynienia ze zbiorem, w którym poszczególne klasy są zgrupowane w przestrzeni cech w mniejszych, zwartych podzbiorach. Podzbiory różnych klas są wzajemnie poprzeplatane. Zdecydowana większość klas miała blisko 1000 próbek.

2. Klasyfikatory

Celem projektu było porównanie jakości klasyfikatorów AdaBoost oraz sieci neuronowej. Dla obu z nich staraliśmy dobrać hiperparametry tak, aby klasyfikator osiągał jak największą dokładność (accuracy).

AdaBoost osiągnął najlepsze wyniki dla 20 estymatorów oraz przy użyciu RandomForest jako klasyfikatora bazowego. Dla sieci NLP próbowaliśmy wielu kombinacji za pomocą algorytmu GridSearch. Ostatecznie wybraliśmy sieć o 5 warstwach, z których każda zawiera po 128 wierzchołków z funkcją aktywacyjną ReLU.

Przy takich ustawieniach MLP i AdaBoost osiągnęły odpowiednio wyniki 92,94% oraz 93,74%.

3. Moc klasyfikatora w zależności od zaszumienia danych wejściowych

Zbadaliśmy, jak zmieniała się dokładność klasyfikacji, gdy w zbiorze treningowym znajdowało się 10%, 20%, 30% próbek z niewłaściwymi etykietami. Dokładność badano na zbiorze testowym.

	noise size	ada boost accuracy	mlp accuracy
0	0.0	0.935279	0.928117
1	0.1	0.935809	0.868170
2	0.2	0.933687	0.837401
3	0.3	0.931034	0.732626

Tab. 2. Dokładność klasyfikacji dla różnych stopni zaszumienia zbioru treningowego

W przypadku MLP jakość klasyfikacji systematycznie spadała. Inaczej było w przypadku klasyfikatora AdaBoost. Zaszumienie rzędu 10% sprawiło, iż jakość klasyfikacji poprawiła się o 0,6%. Klasyfikator ten był także bardziej odporny na większy stopień zaszumienia, uzyskując dokładność 93,1% dla szumu na poziomie 30%.

4. Zmiana jakości klasyfikacji po zastosowaniu PCA

Badaliśmy również wpływ transformaty PCA na jakość klasyfikacji. Przetestowaliśmy dokładność dla transformat z N-D do N-D oraz za N-D do 30-D.

	n_components	ada boost accuracy	mlp accuracy
0	128	0.931034	0.931300
1	30	0.936870	0.937135

Tab. 3. Dokładność klasyfikacji po użyciu transformaty PCA

W obu przypadkach wyniki są nieznacznie lepsze dla 30 komponentów, co pozwala sądzić, że wybór ważniejszych komponentów daje w tym wypadku lepsze wyniki. Generalnie względem bazowej klasyfikacji, po zastosowaniu PCA wyniki niewiele polepszyły się dla sieci neuronowej i pogorszyły dla AdaBoost.

5. Jakość klasyfikatorów na bazie walidacji krzyżowej

Przy pomocy walidacji krzyżowej w obrębie całego zbioru badaliśmy wybrane parametry klasyfikatorów: dokładność, log loss, pole pod krzywą ROC, pole pod krzywą precision-recall, wartość f1, oraz error (w przypadku AdaBoosta). Walidacja opierała się na podziale zbioru na 5 podzbiorów, wyniki w poniższej tabeli to wartości uśrednione.

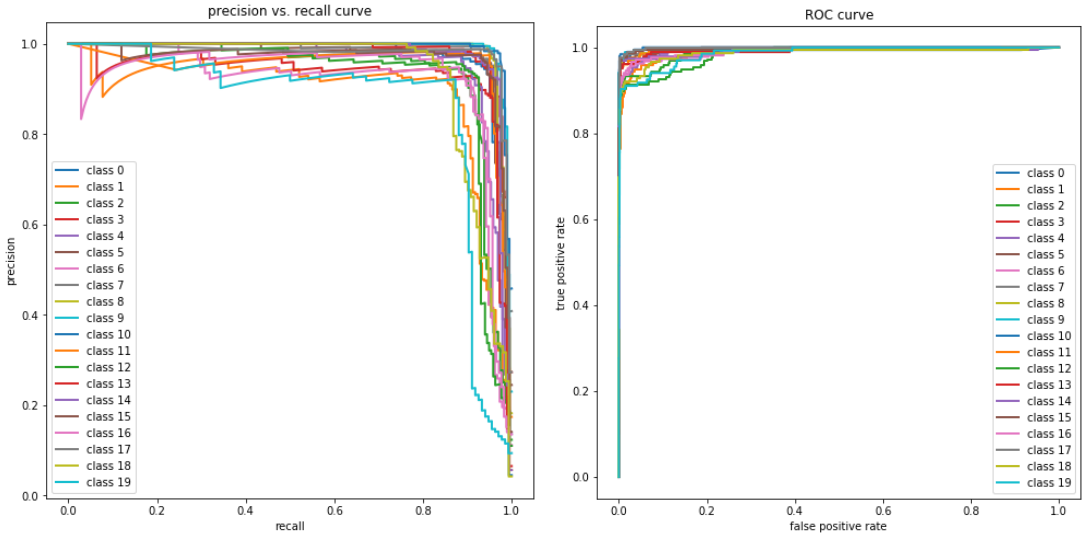
	metric	MLP	AdaBoost
0	accuracy	0.927992	0.938605
1	neg_log_loss	-0.480513	-0.410955
2	roc_auc_ovr_weighted	0.990857	0.992053
3	average_precision	0.946762	0.955265
4	f1_weighted	0.932845	0.939202
5	error	-	0.853313

Tab. 4. Wyniki pomiarów jakości klasyfikatorów uzyskane dzięki walidacji krzyżowej.

Dla każdego badanego parametru AdaBoost okazał się dawać nieznacznie lepsze wyniki od MLP.

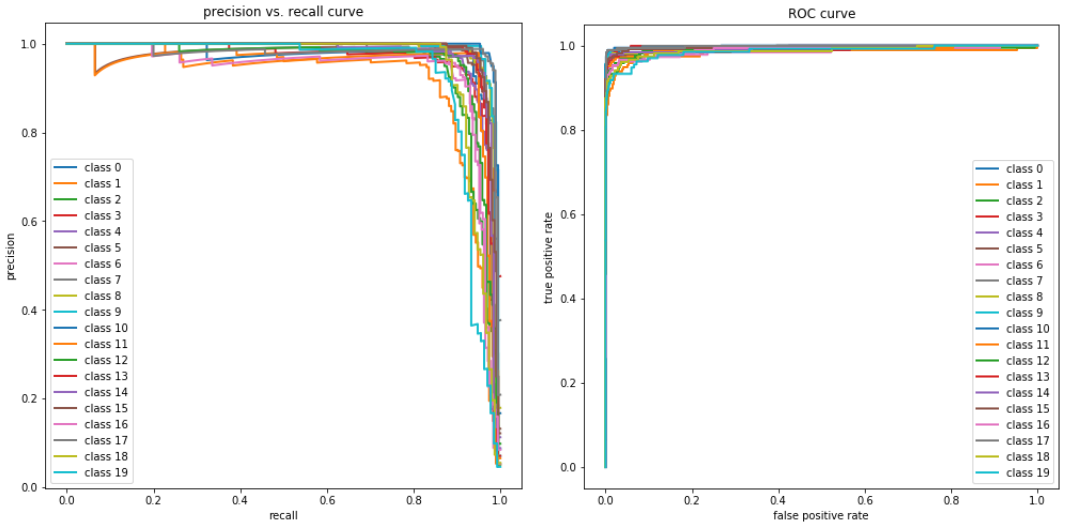
Ponadto sporządziliśmy wykresy krzywych precision-recall oraz ROC. Ponieważ nasze zadanie klasyfikacyjne było niebinarne, zastosowaliśmy klasyfikator OneVsRestClassifier, co pozwoliło sporządzić wykresy krzywych dla wielu klas. Wykresy obrazują wcześniejszy pomiar liczbowy uzyskany drogą walidacji krzyżowej - pole pod krzywymi w przypadku ROC jak i krzywej precision-recall jest bliskie 1, jednak nieznacznie mniejsze dla MLP (zwłaszcza dla krzywej precision-recall).

MLP



Rys. 2. Krzywe precision-recall oraz ROC dla klasyfikatora MLP

AdaBoost



Rys. 3. Krzywe precision-recall oraz ROC dla klasyfikatora AdaBoost

6. Porównanie accuracy i loss dla różnych budżetów czasowych

Porównywaliśmy także jak zmienia się jakość klasyfikacji w zależności od tego jak dużo czasu możemy poświęcić na proces nauczania algorytmów. Dla sieci neuronowych zwiększyliśmy liczbę epok, natomiast dla AdaBoost liczbę komponentów.

	budget	MLP accuracy	MLP log loss	AdaBoost accuracy	AdaBoost log loss
0	short	0.932626	0.376706	0.935544	0.595192
1	medium	0.935279	0.327973	0.937135	0.366630
2	long	0.920690	0.713244	0.935544	0.328208

Tab. 5. Wartości miar klasyfikacji dla różnych budżetów czasowych

Najlepsze wyniki uzyskał średni budżet czasowy. Gdy klasyfikatory były trenowane za długo wtedy doprowadzało to do overfitting'u i jakość klasyfikacji spadała (zarówno funkcja straty jak i dokładność).

7. Zestaw cech głównie odpowiedzialnych za przynależność do klas

W celu dokonania selekcji cech najbardziej odpowiedzialnych za przynależność do poszczególnych klas, zastosowaliśmy klasyfikator RandomForest. Przy jego pomocy określiliśmy wartość współczynnika Giniego dla każdej cechy. Wybraliśmy 20 cech, które uzyskały największą jego wartość.

```
[(32, 0.0233871394569857),
 (56, 0.020319405048144255),
 (71, 0.01976038323134153),
 (15, 0.019610421492664766),
 (18, 0.019595568313208304),
 (97, 0.018089061859414814),
 (112, 0.017840525884859137),
 (78, 0.017495506217595807),
 (114, 0.017063869312689484),
 (40, 0.016827593490835246),
 (116, 0.016723757169138227),
 (51, 0.01671409119683644),
 (60, 0.016565196049268687),
 (16, 0.016494263028860458),
 (55, 0.0161946583593352),
 (99, 0.015678468549301914),
 (27, 0.01558185647368765),
 (81, 0.01502850388874204),
 (110, 0.014740974358385565),
 (33, 0.014626048804849277)]
```

Tab. 6. Wartości współczynnika Giniego dla 20 najbardziej dyskryminujących cech

8. Wnioski

W większości punktów lepsze wyniki osiągnął klasyfikator AdaBoost. Wynika to prawdopodobnie z tego, że dane są wynikiem przetworzenia ich przez sieć neuronową oraz z tego że są dobrze zgrupowane, co ułatwia uzyskiwanie dobrych wyników metodą AdaBoost. Jednak różnice w wynikach pomiędzy tymi algorytmami często są nieznaczne i w obu przypadkach uzyskiwano dokładność powyżej 93%, co jest dobrym wynikiem.