GenderVoice

Drazkowski Hubert, Wilk Marcin

Kwiecień 2021

1 Wstep

Poruszanym przez nas w projekcie problemem było przewidywanie płci na podstawie właściwości akustycznych ich głosu. Na potrzeby stworzenia bazy danych zostały nagrane 3168 próbki dźwiekowe pochodzáce od kobiet lub meżczyzn. Nagrania zostały pretworzone przez pakiety tuneR i seeewave w pakiecie R na przedziale czestotliwości 0hz - 280 hz. Analiza statystyczna tak stworzonych danych została wykonana w programie Python. Dla czytelności i zwiezłości raportu niektóre opisy metod lub rysunki wykorzystywane podczas analizy zostały pominiete. Można je znaleźć w noatatniku jupyter.

Badanie rozpoczyna sie od analizy eksploracyjnej, usuwamy zmienne współliniowe. Na poczatku zmienne deterministycznie wspóliniowe nastepnie zmienne wysoce ze soba skorelowane. Nastepnie dokonujemy przekształceń kilku zmiennych w celu nadania ich rozkładom brzegowym kształtu bardziej zbliżonego do rozkładu normlanego. Kolejnym kroekiem jest poszukiwanie obserwacji odstajacych na dwa sposoby. Pierwszym sposobem jest analiza ich rozkładów brzegowych, drugim algorytm grupujacy DBSCAN.
Nastepnie tworzymy kilka ramek danych odzwierciedlajacych kolejne etapy przetwarzania danych, aby znaleźc odpowiedź jaki wpływ maja te etapy przetwarzania danych na wynik róznych klasyfikatorów. Kończymy nasze badanie aplikujac beterie klasyfikatorów i porównujac ich bład predykcji mierzony odsetkiem poprawnie zaklasyfikowanych przypadków posługujac sie 10 krotna kroswalidacja.

2 Opis bazy danych

Baza danych to 21 zmiennych. Jedna zmienna jest zmienna nominalna wskazujaca na płeć pozostałe 20 zmiennych to zmienne pochodzace z rozkładów ciagłych. Zmienne informuja o właściwościach akustycznych, to jest można wśród nich znaleźć średnia czestotliwość fali, odchylenie standardowe czestotliwości, statystyki spektralne itp. W zbiorze danych nie wystepuja braki danych. Dokładny opis zmiennych można znaleźc miedyinnymi pod adresem https://data.world/ml-

```
RangeIndex: 3168 entries, 0 to 3167
Data columns (total 21 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   meanfreq  3168 non-null   float64
 1   sd        3168 non-null   float64
 2   median    3168 non-null   float64
 3   Q25       3168 non-null   float64
 4   Q75       3168 non-null   float64
 5   IQR       3168 non-null   float64
 6   skew      3168 non-null   float64
 7   kurt      3168 non-null   float64
 8   sp.ent    3168 non-null   float64
 9   sfm       3168 non-null   float64
 10  mode      3168 non-null   float64
 11  centroid  3168 non-null   float64
 12  meanfun   3168 non-null   float64
 13  minfun    3168 non-null   float64
 14  maxfun    3168 non-null   float64
 15  meandom   3168 non-null   float64
 16  mindom    3168 non-null   float64
 17  maxdom    3168 non-null   float64
 18  dfrange   3168 non-null   float64
 19  modindx   3168 non-null   float64
 20  label     3168 non-null   object
```

	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm
count	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000
mean	0.180907	0.057126	0.185621	0.140456	0.224765	0.084309	3.140168	36.568461	0.895127	0.408216
std	0.029918	0.016652	0.036360	0.048680	0.023639	0.042783	4.240529	134.928661	0.044980	0.177521
min	0.039363	0.018363	0.010975	0.000229	0.042946	0.014558	0.141735	2.068455	0.738651	0.036876
25%	0.163662	0.041954	0.169593	0.111087	0.208747	0.042560	1.649569	5.669547	0.861811	0.258041
50%	0.184838	0.059155	0.190032	0.140286	0.225684	0.094280	2.197101	8.318463	0.901767	0.396335
75%	0.199146	0.067020	0.210618	0.175939	0.243660	0.114175	2.931694	13.648905	0.928713	0.533676
max	0.251124	0.115273	0.261224	0.247347	0.273469	0.252225	34.725453	1309.612887	0.981997	0.842936

research/gender-recognition-by-voice. Próba jest zbilansowana, jako że meżczyzni i kobiety sa reprezentowani po równo.

3 Eksploracyjna analiza i przetwarzanie danych

Wiekszość tych zmiennych jest statystykami wygenerowanymi na podstawie nagrań głosów poszczególnych osób, stad musimy uważać na (niekoniecznie unikać) tworzenie statystyk z innych statystyk, takich jak średnia ze średnich. Może to prowadzić do nieścisłości, gdyż nie mamy dostepu do poczatkowych próbek.

Zmienne maja w wiekszości dość małe zakresy, mieszcza sie co najwyżej miedzy 0 a 1, z kilkoma wyjatkami. Duże wartości maksymalne maja również zmienne skew, maxdom i dfrange i kurt.

	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx
count	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000	3168.000000
mean	0.165282	0.180907	0.142807	0.036802	0.258842	0.829211	0.052647	5.047277	4.994630	0.173752
std	0.077203	0.029918	0.032304	0.019220	0.030077	0.525205	0.063299	3.521157	3.520039	0.119454
min	0.000000	0.039363	0.055565	0.009775	0.103093	0.007812	0.004883	0.007812	0.000000	0.000000
25%	0.118016	0.163662	0.116998	0.018223	0.253968	0.419828	0.007812	2.070312	2.044922	0.099766
50%	0.186599	0.184838	0.140519	0.046110	0.271186	0.765795	0.023438	4.992188	4.945312	0.139357
75%	0.221104	0.199146	0.169581	0.047904	0.277457	1.177166	0.070312	7.007812	6.992188	0.209183
max	0.280000	0.251124	0.237636	0.204082	0.279114	2.957682	0.458984	21.867188	21.843750	0.932374

3.1 Współliniowość

W celu zbadania współliniowośći posłużyliśmy sie macierzami korelacji Spearmana i Pearsona oraz współczynnikiem VIF. Z orginalnego zbioru danych zostały usuniete zmienne, które okazały sie być kombinacjami liniowymi pozostałych zmiennych w zbiorze danych. Zarówno macierze korelacji jaki i wyniki VIF obliczeń można znaleźć w notatniku jupyter. Poniżej znajduje sie kika statystyk, które wykluczyliśmy z dalszej analizy.
1. IQR jako, że jest to kombinacja liniowa Q25 i Q75
2. dfrange ktora okazała sie współliniowa z maxdom
3. centroid bedaca wspóliniowa wzgledem meanfreq

Zostały także usuniete wysoce skorelowane ze soba zmienne, jako że potencjalnie przekazywały te sama informacje.
1. skew była mocno skorelowana z kurt
2. sfm jest najsilniej skorelowana z sp.ent oraz ma pozostałe korelacje dużo wieksze od sp.ent, zatem możemy sie jej pozbyć z naszego zbioru.
3. Zmienna meanfreq jest z kolei bardzo silnie skorelowana z median i Q25. Teoretycznie statystyki te nie musza być w ogóle ze soba powiazane, lecz w naszym przypadku bardzo możliwe jest, że meanfreq jest kombinacja liniowa tych dwóch zmiennych i możemy usunać ja z modelu.
4. Niespodzianka jest silna korelacja ujemna wystepuje pomiedzy sd a Q25, jednak przy usunieciu meanfreq oraz centroid musimy zostawić w naszym zbiorze Q25, a możemy usunać sd.

3.2 Przekształcenia zmiennych

Dokonaliśmy kilku przekształceń zmiennych, by były bardziej podobne do rozkładów normalnych. Takimi przkesztalceniami objeliśmy zmienna kurt, maxdom, meandom

$$kurt := log(log(kurt)),$$

$$maxdom := log(maxdom + 1).$$

$$meandom := \sqrt{meandom}$$

Dodatkowo ze wzgledu na fakt, że w przypadku zmiennej mode jej rozkład dla obu płci ma zupełnie różne maksima, kształty oraz punkty koło których gromadziło sie najwiecej obserwacji. Stworzyliśmy zmienna nominalna indykatorowa według reguły

$$mode\_trans := \mathbb{1}_{(mode \in [0.08, 0.15])}(mode)$$

3.3 Obserwacje potencjalnie odstajace

Okazuje sie, że dobrym przekształceniem dla zmiennej meandom jest pierwiastek, wtedy rozkład ten staje sie niemal symetryczny i przypomina nawet w pewnym sensie krzywa Gaussa. Martwi nas jedynie duży słupek w okolicach zera. Sprawdziliśmy czy nie jest to jedna wartość co wskazywałoby na to, że dane zostały w pewien sposób uciete do tej wartości (np. była to minimalna czułość przyrzadów pomiarowych). Aż 61 wartości przyjmuje 0.088388. Drugi najliczniejszym wynikiem jest 0.265165 z 4 reprezentatantami tej wartości. Podobne zjawisko pojawia sie w przypadku zmiennej modindx. Tam 0 jest reprezentowane 65 razy. W obu przypadkach w proporcjach 1:3 sa to głosy kobiet i meżczyzn.

W naszym zbiorze mamy sporo skośnych rozkładów, wiec tam skrajne obserwacje wcale nie musza być odstajace, wynikaja one raczej ze specyfiki rozkładów. Jedynymi pozostałymi zmiennymi gdzie możnaby zauważyć jakieś obserwacje odstajace sa zmienne Q75 oraz kurt. Q75 ma bardzo cienki ogon z lewej strony, za to brak go z prawej, dodatkowo jest to rozkład mocno skupiony na krótkim odcinku, stad nasze podejrzenie. kurt ma z kolei bardzo dziwna "górke" w histogramie z prawej strony, zupełnie nie pasuje ona do rozkładu. Sprawdźmy czy rozkłady te sa w jakiś sposób normalne testem Shapiro-Wilka i czy możemy zastosować regułe 3 sigm. Zarówno Q75 jak i kurt maja bardzo małe p wartości, rzedu 10^-31. Jak widać z p-wartości obu testów nie mamy żadnego prawa uznać je za normalne.

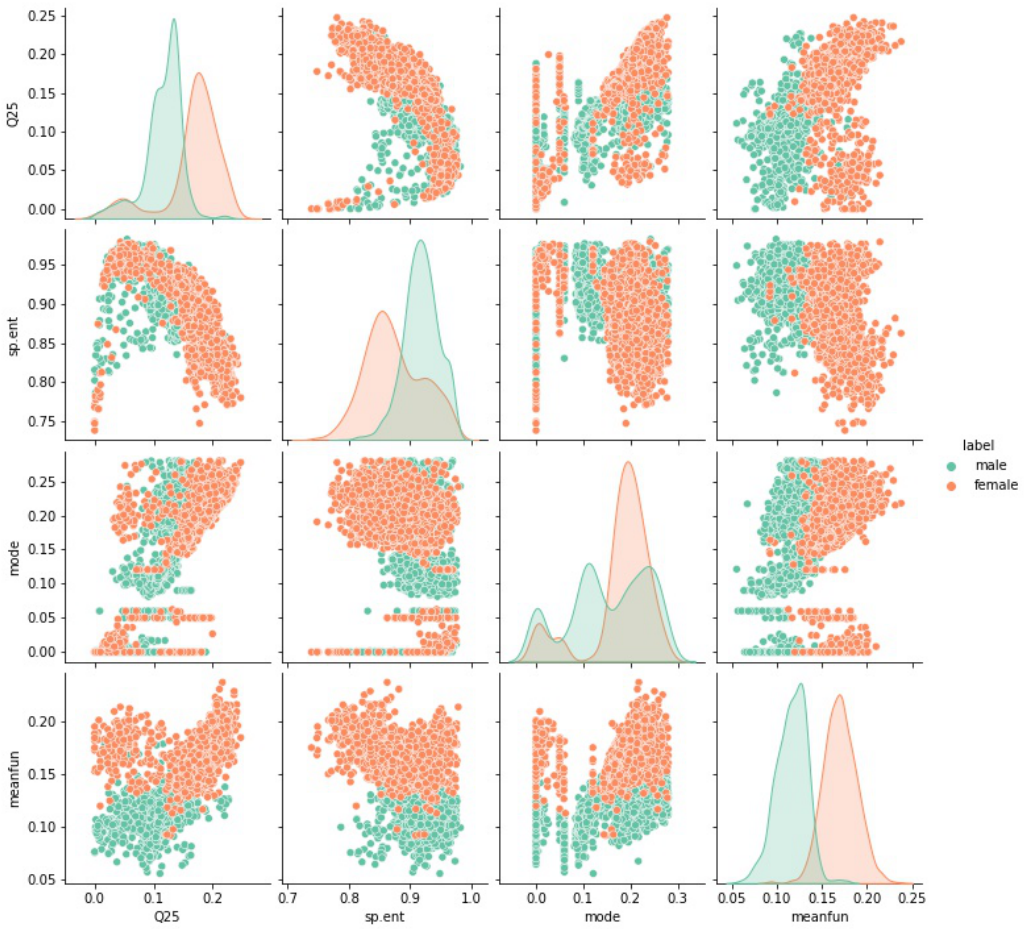
Na nastepnym etapie posłużymy sie nierównościa Markowa. Z tego powodu wycentrowaliśmy najpierw te zmienne. Chcemy sprawdzić czy odstajace sa obserwacje mniejsze od -0.07 dla Q75 oraz wieksze od 1 dla kurt. Z nierówności Markowa wiemy, że

$$\mathbb{P}(|X| > eps) \leq \frac{\mathbb{E}(X^2)}{eps^2} = \frac{Var(X)}{eps^2} = p_X$$

Wartości dla Q75 i kurt to odpowiednio 0.114 oraz 0.129. Zatem tak sie szacuja prawdopodobieństwa, że wartości bezwzgledne wychodza poza te wartości. Nie sa one może rzedu 0.05, należy jednak pamietać, że jest to zazwyczaj dość grube oszacowanie i sporo zawyża faktyczna wartość prawdobodobieństwa.

Posługujac sie kryteriami z przykładu nierówności Markowa, z analizy ucietych wartości meandom i modinx uzyskaliśmy 121 obserwacji potencjalnie odstajacych. Detale ujete sa w notatniku jupyter.

W poszukiwaniu ciekawych algorytmów wyróżniania obserwacji odstajacych natrafiliśmy na algorytm DBSCAN. Nie przyjmuje on założeń na temat rozkładu z



którego pochodza zmienne. Wyróżnia on jedynie obserwacje znacznie oddalone od reszty przy założeniu pewnej metryki mierzacej odległość punktów w wielu wymiarach. Algorytm dokonuje klastrowania punktów do siebie zbliżonych i wyróżnia te obserwacje, które nie zostaja przyporzadkowane do żadnej z klas, które sa od pozostałych punktów znacznie oddalone. Dzieki algorytmowi również zostało wyróżnionych 121 obserwacji odstajacych. Jest to natomiast zbieg okoliczności ponieważ tylko 3 obserwacje pokrywaja sie pomiedzy zbiorami obserwacji odstajacych.

3.4 Ważne zmienne

Niektóre zmienne wydaja sie bardzo dobrze separować klasy kobieta i meżczyzna. Sa to zmienne meanfun, Q25, sp.ent. Ciekawa zależność widać tutaj w przy-

padku zmiennej mode. Rozkłady warunkowe kobiet i mezczyzn w tym przypadku różnia sie tym, że w zakresie zmiennej mode od 0.08 a 0.15 kobiece głosy pojawiaja sie bardzo sporadycznie, zaś jeżeli chodzi o rozkład mezczyzn to osiagana jest w tym przedziale jego moda.

4 Modelowanie

4.1 Opis badania

Stworzyliśmy kilka wersji ramek danych. Zobaczymy czy element preprocessingu dnaych istotnie wpływa na jakość klasyfikacji róznych algorytmów, jeżeli tak, to na które algorytmy.

Nazwa ramki	Operacje
df1	surowy zbiór danych z usunietymi deterministycznie współliniowymi zmiennymi
df2	df1 z dodatkowo usunietymi wysoce skorelowanymi zmiennymi
df3	df2 z przekształcenimi w celu normalizowania
df4	df3 z usunietymi obserwacjami odstajacymi na podstawie analizy rozkładów brzegowych
df5	df3 z usunietymi obserwacjami na podstawie DBSCAN

Druga cześcia badania jest porównanie jakosci klasyfikatorów. Porównujemy ze soba kilka prostych modeli regresji logitowej jako modele benchmarkow. Nastepnie model logitowy z reguaryzacja, lasy losowe, SVM, XGBoost, oraz komitet modeli QDA, regresji logistycznej i naiwnego klasyfikatora bayesa. Porównanie przebiega za pomoca prostego zliczania odsetka poprawnych wskazań w 10 krotnej kroswalidacji.

4.2 Wyniki badania

Po pierwsze zauważyliśmy, że transformacja rozkładów w celu ich znormalizowania nie wpłyneła na jakość klasyfikacji używanych przez nas algorytmów. Odkryliśmy także, że wystandaryzowanie zmiennych polepsza jakość klasyfikacji. W tabeli poniżej przedstawiamy nazwe modelu i średnie accuracy dla df1, df2, df4 i df5. Pierwsza tabela zawiera wyniki przed standaryzacja kolumn druga tabela zawiera wyniki po standaryzacji kolumn. Wyniki to mediana accuracy. Wiecej statystyk i dokładniejsza analiza znajduje sie w notebooku. Przed wystandaryzowaniem kolumn

Nazwa modelu	df1	df2	df4	df5
Regresja logistyczna sam intercept	0.498423	0.498423	0.510672	0.511475
Regresja logistyczna sam meanfun	0.954259	0.954259	0.958941	0.957302
Regresja logstyczna z kara l2	0.894147	0.908373	0.880118	0.875210
Lasy losowe	0.971609	0.966877	0.975410	0.972131
Komitet	0.958991	0.954259	0.970492	0.970492
SVM	0.802215	0.958936	0.960591	0.963881
XGBoost	0.973186	0.973186	0.975383	0.977022

Po wystandaryzowaniu kolumn

Nazwa modelu	df1	df2	df4	df5
Regresja logstyczna z kara l2	0.970032	0.966877	0.980328	0.978689
Lasy losowe	0.971609	0.966877	0.975410	0.972131
Komitet	0.966877	0.965300	0.980296	0.977017
SVM	0.976341	0.962145	0.981946	0.975388
XGBoost	0.973186	0.973186	0.977022	0.975383

5 Wnioski

Standaryzowanie zmiennych wpłynęło w nastepujacy sposób na jakość klasyfikacji prszczególnych modeli:
1. Regresja logistyczna znaczaco zyskała na wystandaryzowaniu zmiennych. Mediana przed zabiegiem wynosiła na poszczególnych
zbiorach od [88%, 91%] teraz jest na poziomie [97%,99%]
2. Las losowy w żaden sposób nie zyskał na standaryzacji
3. Mediana jakości klasyfikacji dla komitetu modeli wzrosła od [95% ; 97%] do [96% ; 97%]
4. Mediana jakości klasyfikacji dla SVM po wszystkich ramkach danych wzrosła od [80%;96%] do [97%;98%]
5. Dla XGBoosta mediana jakości klasyfikacji nie zmieniła sie znaczaco.

Wydaje sie, że prosty model z samym meanfun jest potężny osiagajac niebagatelne 95% jakości stabilnie na wszystkich ramkach danych.

Standaryzacja kolumn nie zmienia wynikow dla lasów losowych, xgboosta. Nieznacznie wpływa na klasyfikacje komitetów, natomiast ma kolsoalne znaczenie w przypakdu SVM i regresji logistycznej z kara RIDGE.

Jakie znaczenie ma pozostały preprocessing ?

Srednio usuwanie obserwacji odstajacych na podstawie ich rozkładów brzegowych radzi sobie lepiej niż wielowymiarowy DBSCAN. Usuwanie obserwacji odstajacych pozwala także na poprawe klasyfikacji wzgledem zbioru tego przekształcenia pozbawionego, ale nie zawsze znaczaco. Usuwanie zmiennych silnie

skorelowanych, ale nie idealnie współlinowych pogarsza jakość klasyfikacji.

Najlepiej klasyfikuje mniej wiecej 1. Regresja logistyczna z regularyzacji 2.SVM 3. XGBoost 4. Komitet 5. Lasy losowe . Aczkowliek nie jest to żadna ścisła dominacja. Sa przypadki ramek danych dla których taka kolejność mogłaby być mylaca