Disfonía en enfermos de Parkinson

Rubén Ibarrondo López y Miren Hayet Otero

15/5/2021

Índice

1 Objetivo

El objetivo principal de este trabajo consiste en encontrar un modelo de clasificación capaz de diferenciar a enfermos de Parkinson de pacientes sanos, en base a registros de voz. Para ello, primero se van a analizar las características de los datos de los que se dispone, y después se ajustarán y compararán diferentes técnicas de clasificación.

2 Análisis preliminar

Es necesario realizar un análisis preliminar de los datos para después obtener un modelo lo más fácil de interpretar y mejor posible.

La base de datos utilizada se puede consultar aquí. Se dispone de 195 registros de voz correspondientes a 31 pacientes, de los cuales hay 23 enfermos de Parkinson. Para cada registro se han recogido 23 medidas relacionadas con la voz:

- MDVP.Fo.Hz: Frecuencia vocal fundamental media.
- MDVP.Fhi.Hz : Frecuencia vocal fundamental máxima.
- MDVP.Flo.Hz: Frecuencia vocal fundamental mínima.
- MDVP.Jitter, MDVP.Jitter.Abs, MDVP.RAP, MDVP.PPQ, Jitter.DDP: Medidas de variación en la frecuencia fundamental.
- MDVP.Shimmer, MDVP.Shimmer.dB, Shimmer.APQ3, Shimmer.APQ5, MDVP.APQ, Shimmer.DDA: Medidas de variación en la amplitud.
- NHR,HNR: Medidas del ratio entre el ruido y las componentes tonales de la voz.
- status: Estado de salud del paciente. 1-Parkinson, 0-Sano.
- RPDE, D2: Medidas no-lineales de complejidad dinámica.
- DFA: Exponente escalador de fractal de señal.
- spread1, spread2, PPE: Medidas no-lineales de la variación de la frecuencia fundamental.

En este caso no hay ningún dato ausente por lo que no va a ser necesaria ninguna estrategia de imputación.

A continuación se puede ver un resumen de las diferentes variables:

```
##      name            MDVP.Fo.Hz.      MDVP.Fhi.Hz.    MDVP.Flo.Hz.
##  Length:195         Min.   : 88.33   Min.   :102.1   Min.   : 65.48
##  Class :character   1st Qu.:117.57   1st Qu.:134.9   1st Qu.: 84.29
##  Mode  :character   Median :148.79   Median :175.8   Median :104.31
##                     Mean   :154.23   Mean   :197.1   Mean   :116.32
##                     3rd Qu.:182.77   3rd Qu.:224.2   3rd Qu.:140.02
##                     Max.   :260.11   Max.   :592.0   Max.   :239.17
##  MDVP.Jitter...     MDVP.Jitter.Abs.       MDVP.RAP           MDVP.PPQ
##  Min.   :0.001680   Min.   :7.000e-06   Min.   :0.000680   Min.   :0.000920
##  1st Qu.:0.003460   1st Qu.:2.000e-05   1st Qu.:0.001660   1st Qu.:0.001860
##  Median :0.004940   Median :3.000e-05   Median :0.002500   Median :0.002690
##  Mean   :0.006220   Mean   :4.396e-05   Mean   :0.003306   Mean   :0.003446
##  3rd Qu.:0.007365   3rd Qu.:6.000e-05   3rd Qu.:0.003835   3rd Qu.:0.003955
```

```
##  Max.   :0.033160   Max.   :2.600e-04   Max.   :0.021440   Max.   :0.019580
##    Jitter.DDP        MDVP.Shimmer     MDVP.Shimmer.dB.  Shimmer.APQ3
##  Min.   :0.002040   Min.   :0.00954   Min.   :0.0850   Min.   :0.004550
##  1st Qu.:0.004985   1st Qu.:0.01650   1st Qu.:0.1485   1st Qu.:0.008245
##  Median :0.007490   Median :0.02297   Median :0.2210   Median :0.012790
##  Mean   :0.009920   Mean   :0.02971   Mean   :0.2823   Mean   :0.015664
##  3rd Qu.:0.011505   3rd Qu.:0.03789   3rd Qu.:0.3500   3rd Qu.:0.020265
##  Max.   :0.064330   Max.   :0.11908   Max.   :1.3020   Max.   :0.056470
##   Shimmer.APQ5        MDVP.APQ        Shimmer.DDA           NHR
##  Min.   :0.00570   Min.   :0.00719   Min.   :0.01364   Min.   :0.000650
##  1st Qu.:0.00958   1st Qu.:0.01308   1st Qu.:0.02474   1st Qu.:0.005925
##  Median :0.01347   Median :0.01826   Median :0.03836   Median :0.011660
##  Mean   :0.01788   Mean   :0.02408   Mean   :0.04699   Mean   :0.024847
##  3rd Qu.:0.02238   3rd Qu.:0.02940   3rd Qu.:0.06080   3rd Qu.:0.025640
##  Max.   :0.07940   Max.   :0.13778   Max.   :0.16942   Max.   :0.314820
##       HNR             status            RPDE             DFA
##  Min.   : 8.441   Min.   :0.0000   Min.   :0.2566   Min.   :0.5743
##  1st Qu.:19.198   1st Qu.:1.0000   1st Qu.:0.4213   1st Qu.:0.6748
##  Median :22.085   Median :1.0000   Median :0.4960   Median :0.7223
##  Mean   :21.886   Mean   :0.7538   Mean   :0.4985   Mean   :0.7181
##  3rd Qu.:25.076   3rd Qu.:1.0000   3rd Qu.:0.5876   3rd Qu.:0.7619
##  Max.   :33.047   Max.   :1.0000   Max.   :0.6852   Max.   :0.8253
##     spread1          spread2               D2              PPE
##  Min.   :-7.965   Min.   :0.006274   Min.   :1.423   Min.   :0.04454
##  1st Qu.:-6.450   1st Qu.:0.174350   1st Qu.:2.099   1st Qu.:0.13745
##  Median :-5.721   Median :0.218885   Median :2.362   Median :0.19405
##  Mean   :-5.684   Mean   :0.226510   Mean   :2.382   Mean   :0.20655
##  3rd Qu.:-5.046   3rd Qu.:0.279234   3rd Qu.:2.636   3rd Qu.:0.25298
##  Max.   :-2.434   Max.   :0.450493   Max.   :3.671   Max.   :0.52737
```

No parece haber ningún dato disparatado por lo que parecen ser variables coherentes O A LO MEJOR NO Y COMPROBAR COHERENCIA DATOS INUSUALES!! BOXPLOT????????. La distribución dela variable de clasificación nos indica que en torno a un 75% de los registros corresponden a enfermos de Parkinson.

A la hora de crear cualquier modelo de clasificación es importante que la cantidad de variabables que lo forman sea lo menor posible, ya que esto facilita su aplicación e interpretación. Muchas veces las medidas/variables de las que se dispone no suelen aportar demasiada información a la hora de clasificar, ya sea por que no están relacionadas con la variable de clasificación o porque no presentan gran variabilidad. También puede ocurrir que algunas variables estén altamente correladas entre sí, por lo que si se incluyen todas en el modelo, no van a aportar nueva información a la hora de clasificar.

Comencemos por ver si hay alguna variable con poca variabilidad:

```
##                  freqRatio percentUnique zeroVar   nzv
## MDVP.Fo.Hz.       1.000000     100.00000   FALSE FALSE
## MDVP.Fhi.Hz.      1.000000     100.00000   FALSE FALSE
## MDVP.Flo.Hz.      1.000000     100.00000   FALSE FALSE
## MDVP.Jitter...    1.000000      88.71795   FALSE FALSE
## MDVP.Jitter.Abs.  1.642857       9.74359   FALSE FALSE
## MDVP.RAP          1.666667      79.48718   FALSE FALSE
## MDVP.PPQ          1.333333      84.61538   FALSE FALSE
## Jitter.DDP        1.500000      92.30769   FALSE FALSE
## MDVP.Shimmer      1.000000      96.41026   FALSE FALSE
## MDVP.Shimmer.dB.  1.250000      76.41026   FALSE FALSE
## Shimmer.APQ3      1.000000      94.35897   FALSE FALSE
## Shimmer.APQ5      1.000000      96.92308   FALSE FALSE
## MDVP.APQ          1.000000      96.92308   FALSE FALSE
## Shimmer.DDA       1.000000      96.92308   FALSE FALSE
## NHR               1.000000      94.87179   FALSE FALSE
```

```
## HNR               1.000000     100.00000   FALSE FALSE
## RPDE              1.000000     100.00000   FALSE FALSE
## DFA               1.000000     100.00000   FALSE FALSE
## spread1           1.000000     100.00000   FALSE FALSE
## spread2           2.000000      99.48718   FALSE FALSE
## D2                1.000000     100.00000   FALSE FALSE
## PPE               1.000000     100.00000   FALSE FALSE
```

Todas las variables presentan una variabilidad suficiente como para poder aportar información en la clasificación.

Para comprobar su Veamos que importancia tiene cada variable en relación con la variable de clasificación:

```
##             DFA     MDVP.Fhi.Hz.     MDVP.Flo.Hz.      MDVP.Fo.Hz.
##       0.6498016        0.6748866        0.6972789        0.7006803
##            RPDE               D2              HNR      Shimmer.DDA
##       0.7071995        0.7249150        0.7379535        0.7546769
##    Shimmer.APQ3     Shimmer.APQ5              NHR         MDVP.RAP
##       0.7548186        0.7699121        0.7731718        0.7769274
##      Jitter.DDP   MDVP.Jitter...     MDVP.Shimmer MDVP.Shimmer.dB.
##       0.7774235        0.7777069        0.7827381        0.7850765
##        MDVP.PPQ MDVP.Jitter.Abs.          spread2         MDVP.APQ
##       0.7872024        0.7889739        0.8136338        0.8258929
##         spread1              PPE
##       0.8969671        0.8969671
```

Ninguna variable obtiene una puntuación que nos asegure que no es lo suficientemente importante como para no incluirla en el modelo.

Por último, nos queda comprobar si existe correlación entre las variables. En el gráfico @ref{fig:corr} se pueden las correlaciones más altas entre variables. Concretamente se distinguen en 4 tonalidades que van desde el azul oscuro al claro las correlaciones mayores a 0.95, 0.9, 0.85 y 0.8 respectivamente.

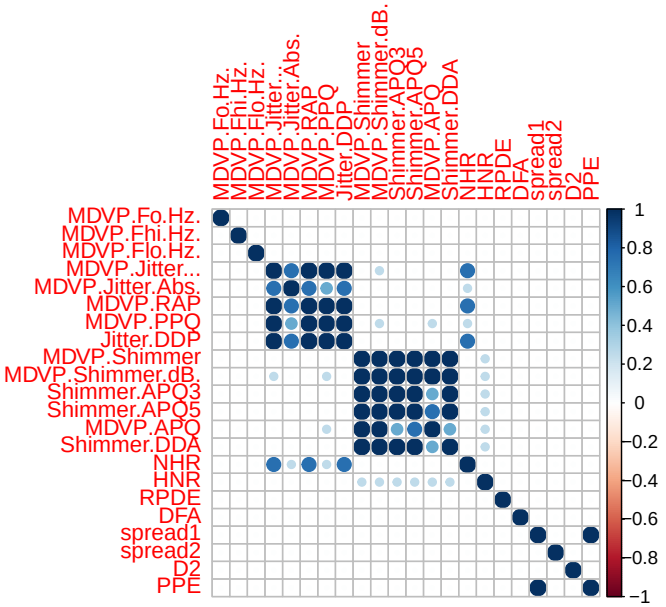

Figure 1: Correlación entre variables

Por lo tanto, si establecemos 0.95 como la máxima correlación que pueden tener dos variables en el modelo, tendremos que escoger una variable entre MDVP.Jitter, MDVP.RAP, MDVP.PPQ y Jitter.DDP, entre MDVP.Shimmer y MDVP.Shimmer.dB y entre spread1 y PPE. Basándonos en la importancia de las variables nos quedaremos con MVDP.PPQ, MDVP.Shimmer.dB y spread1.