Informe árboles de decisión

Diego Hernández Jiménez

1. Introducción

Este trabajo constituye una continuación de un proyecto anterior en el que se pretendía construir un modelo capaz de clasificar de manera precisa a una muestra de sujetos en dos clases, "individuos que han sufrido ictus" (*Yes*) e "individuos que no han sufrido ictus" (*No*). Se mantiene aquí el mismo objetivo y se trabaja con la misma base de datos. En aquella ocasión se empleó el análsis discriminante como técnica de clasificación, pero pudo observarse que presentaba algunos inconvenientes. En primer lugar, requiere predictores cuantitativos, lo cual obligaba a descartar muchos atributos potencialmente útiles. En segundo lugar, se basa en unos supuestos estadísticos exigentes, que resultan difíciles de cumplir.

Por ello, en esta ocasión se opta por un método alternativo. Se propone el uso de clasificadores basados en árboles de decisión. Estos clasificadores permiten el uso de variables categóricas, o variables cuantitativas que han sido discretizadas. Otra ventaja es que no requieren asunciones acerca de las distribuciones de las clases (Alpaydin, 2020).

Siguiendo a (Boehme et al., 2017), se consideran como variables predictoras más relevantes: *age_cat*, *hypertension*, *glucose_cat*, *heart_disease*, *gender*,*bmi_cat*. *age_cat* es una variable categórica creada a partir de *age* con los niveles *Kid* (edad menor o igual a 14 años), *Young* (edad comprendida entre 14 y 25), *Mature* (edad comprendida entre 25 y 64), *Elder* (edad mayor de 64).*hypertension* y *heart_disease* son variables dicotómicas que indican presencia o ausencia de hipertensión y ¿enfermedades cardiovasculares?, respectivamente. *glucose_cat* se crea a partir de *avg_glucose_level* y los niveles son *Diabetes* (nivel de glucosa mayor a 125 mg/dl), *Pre-diabetes* (nivel de glucosa entre 100 y 125), *Normal* (nivel de glucosa inferior o igual a 100). La partición se hace siguiendo los criterios diagnósticos recomendados (Organización Mundial de la Salud, 2006).El atributo *gender* contiene tres niveles en los datos originales, *male*, *female*, y *other*. Aquí se ha recodificado como una variable dicotómica y los 11 casos de la categoría *other*, se deciden considerar como valores perdidos. Por último, *bmi_cat* posee las categorías *Obese* (*bmi* mayor o igual 30), *Overweight* (bmi entre 25 y 30), *Normal* (bmi entre 18.5 y 25) y *Underweight* (bmi menor de 18.5). Los puntos de corte están basados en los proporcionados por la Organización Mundial de la Salud y no se hace distinción por grupo de edad. No se tratan los valores perdidos.

2. Métodos y herramientas

Un problema que presenta la base de datos es el del gran desequilibrio en las clases de la variable a predecir. En un trabajo anterior se recurrió a métodos basados en el submuestreo a partir de la clase mayoritaria. En este trabajo se adopta un enfoque diferente, pero que puede resultar incluso más apropiado, un enfoque basado en el aprendizaje sensible a los costes (He y Garcia, 2009). Se mantiene la muestra original, pero se introduce en todos los modelos una matriz de costes que en este caso penaliza los falsos negativos. Igualmente, dado que en condiciones de desequilibrio de clases la precisión resulta un estadístico poco informativo ("paradoja de la precisión"), se toman como principales medidas del rendimiento del modelo la curva ROC y el área bajo la curva (AUC), así como el valor F. No resulta sencillo llevar a cabo este enfoque mediante Rapidminerm aunque posee el operador Metacost, por lo que solo se utiliza Matlab. El ajuste y evaluación de modelos se raliza en tres etapas.

1. En una primera etapa se ajustan y comparan los primeros modelos para dterminar la matriz de costes más adecuada y para hacer una selección de variables. Para ello se adopta la estrategia de dividir la muestra completa en conjunto de entrenamiento y conjunto de validación. Debido al desequilibrio de clases, este muestreo es estratificado, para asegurar el mantenimiento de las probabilidades a priori en los dos conjuntos de datos. Todos los modelos se ajustan en la misma muestra y se evalúan en la restante para poder ser comparados. Un primer modelo, "modelo base", consistente en un árbol de decisión con todas las variables incluidas y parámetros por defecto (salvo el criterio de división de los nodos, que está basado en la ganancia de información) se utiliza como referencia. El código de este primer árbol es el siguiente:

```
% stroke -> dataset
rng(13)
allfeatures=stroke(:,["age_cat","hypertension","glucose_cat", ...
    "heart_disease","gender","bmi_cat"]);
group=stroke{:,"stroke"};

% split in train and test set
part=cvpartition(group,'HoldOut',.3, 'Stratify',true); % stratify to make sure same priors
ids_train=training(part);
ids_test=test(part);

% Split criterion based on information gain, no cost for false negatives
treebase=fitctree(allfeatures(ids_train,:),group(ids_train,:), ...
    'CategoricalPredictors','all', ...
    'SplitCriterion','deviance');

% predictions and posterior probabilities...........
[predsbase, postprobsbase]=predict(treebase,features(ids_test,:));
% model...........
view(treebase)
treebase.ModelParameters
% assessment...........
confusionchart(group(ids_test,:),predsbase);
c=confusionmatStats(group(ids_test,:),predsbase); % special function
accuracy=c.accuracy; sensitivity=c.sensitivity; specificity=c.specificity; Fscore=c.Fscore;
table(accuracy,sensitivity,specificity,Fscore,'RowNames',{'Class No','Class Yes'})

% ROC...........
[FPRbase, TPRbase, ~, AUCbase]=perfcurve(group(ids_test),postprobsbase(:,2),'Yes');
area(FPRbase,TPRbase,'FaceColor',[0.8 0.89 0.95]);
grid on
text(.5,.6,['AUC = ' num2str(AUCbase)],"FontSize",16);
xlabel('False positive rate')
ylabel('True positive rate')
title('ROC Curve for base Decision Tree')

% important features...........
imp=predictorImportance(treebase);
figure;
bar(imp);
title('Predictor Importance Estimates');
ylabel('Estimates');
xlabel('Predictors');
h = gca;
h.XTickLabel = treebase.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```

La salida proporciona la matriz de confusión, una tabla con los estadísticos de precisión, sensibilidad, especificidad y valor F_1, un gráfico con la curva ROC y AUC asociada y un diagrama de barras con la importancia de cada variable. Este proceso se utiliza con cada modelo. En función de los resultados observados se van añadiendo modificaciones.

2. En una segunda parte, habiendo hecho una selección preliminar de predictores y de matriz de costes

se procede al ajuste de hiperparámetros. En principio se opta por el uso de optimización bayesiana para este proceso, pero dado que un objetivo adicional del trabajo es el de proponer un modelo de poca complejidad, se permite la modificación manual de hiperparámetros referidos a la profundidad del árbol si el modelo optimizado resulta aún excesivamente complejo.

3. Este modelo optimizado se evalúa mediante validación cruzada para obtener una mejor estimación del error de generalización, ya que hasta este punto todas las evaluaciones se han hecho a partir de una única partición de la muestra.

3. Resultados y discusión

Como se esperaba, el modelo base obtiene una precisión del 98.2%, pero una sensibilidad y un valor F de cero. Su valor AUC de 0.5 refleja su inadecuación. Obviamente, ningún predictor resulta relevante para el modelo. El primer modelo es equivalente al base, pero tiene una matriz de costes asociada de `[0 10;100 0]` , es decir, penaliza gravemente las predicciones *No* erróneas. La sensibilidad y el valor F_1 mejoran, obteniéndose valor de 22.13% y 0.1308 respectivamente. Una mejor sustantiva se da en el valor del AUC, que es de 0.808 aproximadamente. Se observa en el diagrama de barras que la variable *gender* es la menos importante, por lo que se deshecha???. El segundo modelo a probar tiene una matriz de costes que penaliza menos `[0 1;70 0]` (los valores realmente se han ido ajustando por ensayo y error). No está presente la variable *gender*. A pesar de ello tiene buen equilibrio entre sensibilidad (80%) y especificidad (72.08%), con un valor F_1 de 0.09. El AUC es de 0.815. No se aprecia que la variable *glucose_cat* sea un buen predictor, por lo que se decide eliminarlo. *hypertension* tampoco parece ser un buen predictor, pero se decide mantener por razones teóricas y porque esa variable también está muy desequilibrada en la muestra, con lo que es posible que la partición inicial haya hecho una repartición sesgada de casos (muestra de entrenamiento con pocos casos *hypertension*=*Yes*). El modelo 3 contiene ahora 4 variables y la misma matriz de costes. Los valores de los estadísticos son muy similares, aunque el AUC se incrementa ligeramente a 0.823 aproximadamente, aunque no se evalúa si la diferencia con respecto al modelo anterior es estadísticamente significativa. Ninguna de las variables queda descartada.

En los tres modelos se observa, como se esperaba, que la variable *age_cat* tiene un peso muy superior en el modelo frente a las demás variables. Por ese motivo, se decide ajustar un último modelo que incluye todos los predictores salvo *age_cat*. Con ello se pretende comparar el peso de las distintas variables cuando no se tiene en cuenta el grupo de edad. El diagrama de barras (Figure 1) permite llegar a las mismas conclusiones que anteriormente. *gender* y *glucose_cat* son las variables menos relevantes, por lo que se mantiene la decisión de mantenerlas fuera del modelo final.

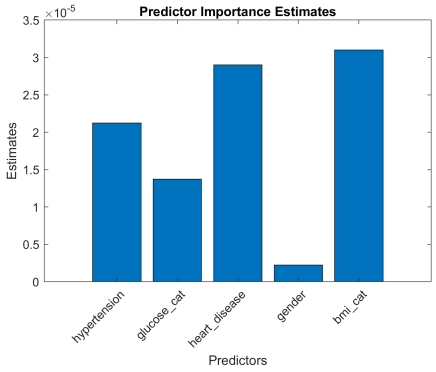

Figure 1: Importancia de los predictores

Una comparación visual de las curvas ROC pone de manifiesto que el modelo 3, es el más adecuado.

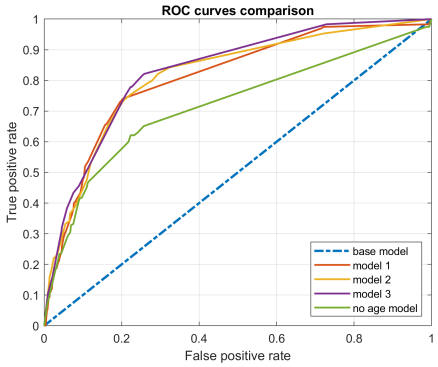

Figure 2: Comparación de curvas ROC

En la etapa 2 se repite el proceso de ajuste con las variables y matriz de costes del modelo 3, pero se realiza una optimización de los hiperparámetros. Este modelo tiene una sensibilidad de 81.7%, una especificidad de 74.56%, un valor F de 0.10 y una AUC de 0.83. Los parámetros del modelo son los siguientes:

```
ans =
        SplitCriterion: 'gdi'
             MinParent: 10
               MinLeaf: 1
             MaxSplits: 29
          NVarToSample: 'all'
           MergeLeaves: 'on'
                 Prune: 'on'
        PruneCriterion: 'error'
               QEToler: []
            NSurrogate: 0
                MaxCat: 10
                AlgCat: 'auto'
    PredictorSelection: 'allsplits'
          UseChisqTest: 1
                Stream: []
          Reproducible: 0
               Version: 2
                Method: 'Tree'
                  Type: 'classification'
```

A pesar de ser resultados optimizados, la complejidad puede reducirse, por ejemplo aumentado el número mínimo de casos en los nodos terminales u hojas. De hecho, si comprobamos los errores de modelos estimados con validación cruzada con diferentes valores para el `MinLeaf` se ve que éste produce mejores resultados (al menos en términos de precisión) con valores entre 200 y 400.

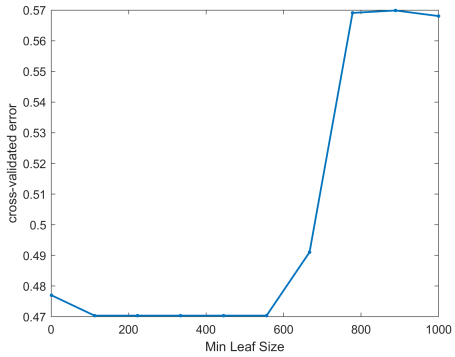

Figure 3: Error de clasificación en función del número mínimo de casos por hoja

Se vuelve ajustar el modelo con un número mínimo de casos por hoja 250 y el resto de parámetros optimizados, esta vez mediante validación cruzada. Los resultados se muestran a continuación.

	Predicción=No	Predicción=Yes
Verdadero=No	31916	10701
Verdadero=Yes	144	639

Los estadísticos de rendimiento son Precisión: 75.01% , Sensibilidad: 81.61%, Especificidad: 74.89% , valor F_1: 0.1054 , AUC: 0.8118

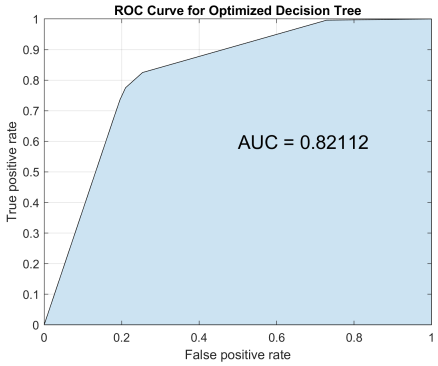

Figure 4: Curva ROC y AUC de modelo optimizado

El modelo de árbol en sí puede resumirse en las siguientes reglas de decisión (previo a la poda había 29 divisiones o *splits*):

```
Decision tree for classification
 1  if age_cat=Elder then node 2 elseif age_cat in {Kid Mature Young} then node 3 else Yes
 2  class = Yes
 3  if age_cat in {Kid Young} then node 4 elseif age_cat=Mature then node 5 else No
 4  class = No
```

```
 5  if bmi_cat=NA then node 6 elseif bmi_cat in {Normal Obese Overweight Underweight}
then node 7 else No
 6  class = Yes
 7  if heart_disease=No then node 8 elseif heart_disease=Yes then node 9 else No
 8  if hypertension=No then node 10 elseif hypertension=Yes then node 11 else No
 9  class = Yes
10  class = No
11  class = Yes
```

4. Conclusiones

La aproximación basada en aprendizaje sensible a costes ha producido buenos resultados en general, aunque los valores de F_1 son reducidos. No obstante, la precisión no se ha resentido tanto como podría haberse esperado, habiéndose hallado un valor similar al que se encontró al emplear funciones discriminantes. La ventaja en esta ocasión, es que podemos tener más confianza en el resultado, pues se ha obtenido a partir de toda la muestra, no como resultado de promediar la precisión de distintos modelos estimados con distintas muestras.

Merece la pena destacar también los predictores que han entrado en el modelo final. Contrariamente a lo que podría esperarse a nivel teórico (Boehme et al., 2017), el hecho de ser o no ser diabético (*glucose_cat*) no ha resultado determinante para realizar predicciones. De nuevo la edad ha sido la variable más "dominante". De hecho, puede observarse en el árbol de decisión que el mero hecho de pertenecer a la categoría *Elder* (más de 64 años) es suficiente para que se prediga que se sufrirá ictus. En futuros trabajos puede ser conveniente considerar de manera específica esta variable, ajustando modelos que la incluyan y otros que no, por ejemplo, para observar mejor el comportamiento del modelo y poder refinarlo más, como se ha hecho en esta ocasión en la etapa 1.

5. Referencias bibliográficas

Alpaydin (2020). Decision Trees. En E. Alpaydin, *Introduction to machine learning* (pp.185). MIT Press.

Boehme, A. K., Esenwa, C., y Elkind, M. S. (2017). Stroke risk factors, genetics, and prevention. *Circulation research*, 120(3), 472-495.

He, H., y Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9), 1263-1284.

Organización Mundial de la Salud, (2006). *Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia. Report of a WHo/IDf Consultation.* Recuperado de https://www.who.int/diabetes/publications/Definition%20and%20diagnosis%20of%20diabetes_new.pdf