Cajamar Datathon 2020

RETO MINSAIT LAND CLASSIFICATION

Equipo Sigma

Víctor Callejas Fuentes

LinkedIn: https://www.linkedin.com/in/victor-callejas-fuentes/
Email: viccalfu@inf.upv.es

25 de marzo de 2020

Resumen

En este documento se recogen las técnicas usadas así como los resultados obtenidos en la resolución de este reto. El código se ha desarrollado de forma que sea fácilmente reproducible, para este fin también se incluye en el repositorio una copia del enviroment. Este documento viene acompañado de referencias externas y internas al código usado en formato de Jupyter Notebooks.

NOTA PARA EL JURADO

Este documento lo he desarrollado para ir organizandome durante la resolución del reto por lo que hay mucha información redundante para vosotros.
Os recomiendo que salteís al apartado 3.3.

Índice

1 Planteamiento del problema

El reto propuesto es el de clasificar el tipo de terreno a partir de datos tabulares obtenidos de imágenes satelitales[7] y datos del catastro del gobierno[8]. La métrica a optimizar:

$$Accuracy = \frac{\text{Registros bien clasificados}}{\text{Número total de registros}}$$

El número de clases posibles son 7:

- RESIDENTIAL
- INDUSTRIAL
- PUBLIC
- RETAIL
- OFFICE
- OTHER
- AGRICULTURE

2 Conocimiento sobre el dominio

Las aclaraciones en este apartado son convenientes para la resolución del reto.

2.1 Satelites

La agencia espacial europea a través del programa Copernicus puso en orbita baja diversos satélites para tomar fotografías de alta resolución de la Tierra. En concreto en este proyecto se utilizan datos recogidos de los satélites Sentinel II(A y B)[10]. Estas son sus características más relevantes:

- Actualización de la zona cada 5 días o menos
- Resolución espacial de 10 a 60 metros
- Datos fotográficos en las bandas visible y infrarroja

En concreto en nuestros datos de las 13 bandas disponibles se encuentran la 2,3,4(azul, verde, roja) y 8(NIR)(Infrarroja, genera un mapa térmico), la resolución espacial de ambas es de 10 metros, lo que quiere decir que un píxel representa $100m^2$ de terreno[11]. El resto de bandas del satélite se presupone que se han descartado porque sus resoluciones espaciales son de 20m o más y no guardan tanta relación.

2.2 Catastro

El Catastro Inmobiliario es un registro administrativo dependiente del Ministerio de Hacienda en el que se describen los bienes inmuebles rústicos, urbanos y de características especiales.

En este registro se encuentran datos sobre las fincas como, localización, geometría, superficie, tipo, año de construcción y calidad...

3 Datos

Los datos proporcionados contienen 55 variables, con 103230 observaciones para el entrenamiento del modelo y 5618 observaciones para la entrega de predicciones.

3.1 Origen de los datos

Estos datos se han obtenido al juntar los datos satelitales con los del catastro, suponemos que se han unido a través de las cordenadas X, Y.

3.2 Variables

- ID

  Guarda relación con Id catastrales pero están ofuscados para evitar Data Leakages.

  Solo sirven para identificar al registro

- Cordenadas

  Son X y Y, en los datos catastrales corresponden con los centroides de las parcelas[9].

  En nuestros datos son la longitud-latitud, pero están modificadas (escaladas y desplazadas aleatoriamente). Guardan relación entre los registros pero no permite su ubicación.

- Imagenes Satelitales

  Valores tomados de 4 bandas: rojo(4), verde(3), azul(2) y infrarroja(8). Todas con resolución espacial de 10m.

  Para cada banda se definen 11 valores, que representa la densidad del color por décil.

- Area

  Superficie en m^2 de la parcela

- Geom

  Idealista no ha proporcionado el significado.

  Solo se conoce que condensan la información geométrica.

  Presuponemos que podría ser el perímetro normalizado, la volumetría o distancia desde el centro de la finca a un punto...

- Construction year

  Año de construcción de parcelas colindantes

- Max. Bulding Floor

  Número del último piso de parcelas colindantes

- Cadastral Quality ID

  Calidad de la construcción de la parcela. Toma valores de 1 a 9, yendo de mejor a peor. Excepcionalmente, puede aparecer como A, B y C para lo mejor que lo mejor[9].

  En esta variable existe una relación entre las distintas clases, aunque no se ha podido

encontrar(lineal, exp...), supondré una relación lineal.

- CLASE

  Variable a predecir.

  Clases muy desbalanceadas:

RESIDENTIAL	87.4 %
INDUSTRIAL	4.4 %
PUBLIC	2.9 %
RETAIL	2 %
OFFICE	1.8 %
OTHER	1.3 %
AGRICULTURE	0.3 %

3.3 Missing values

Hay 40 valores nulos en el train set y 14 valores nulos en el test set. En ambos casos la cantidad de observaciones con valores nulos es exactamente la mitad, 20 y 7 respectivamente, suponen el 0.0005 % de las observaciones en train y 0.0097 % en test. Esto es porque solo hay dos variables que contengan valores nulos y no existe ninguna observación en la que solo una de dichas variables sea nula.

- Max. Bulding Floor
- Cadastral Quality ID

Se observa que predomina la Clase AGRICULTURA en estas observaciones.

3.4 Distribución Modelar y Estimar

Al comprobar las distribuciones de las variables del Modelar set y del Estimar set se observa que vienen de la misma distribución, menos la variable AREA.

Cuadro 1: Distribución área Modelar y Estimar

	Modelar	Estimar	Diferencia
Media	441	967	526
Std	1870	3027	1157
Min	0	1	0
25 %	97	119	21
50 %	172	263	91
75 %	344	727	384
Max	238059	104563	133496

Teniendo en cuenta que el area media de la clase RESIDENTIAL es de 281 y para el resto de clases la media es mayor que 1000. Voy a suponer que en el test set habrán bastantes menos observaciones del tipo RESIDENTIAL.

4 Análisis Exploratorio de los Datos

Tras el análisis hecho en esta sección, se observa que las clases se superponen mucho en todas las dimensiones, por lo que crear nuevas variables en la fase de feature engineering será vital.

Probablemente sea imposible conseguir un modelo con un 100 % de accuracy. Conociendo esto, el desbalanceamiento de las clases y que la distribución de las clases en Estimar es diferente, se enfoca este reto poniendo más énfasis en crear un modelo robusto que prediga bien en el Estiamar set más que conseguir uno con mayor precisión en el Modelar set.

4.1 Imputando los valores nulos

Aunque el modelo elegido en el apartado 6 (Decission Trees) puede lidiar con valores nulos y estos aportar significado a los datos, se ha decido imputarlos por simplicidad ya que estos no suponen ni un 0.01 % de las observaciones y si que es necesario eliminarlos o imputarlos a la hora de balancear los datos y aplicar la técnica SMOTE.

Para ello, como la gran mayoría de las observaciones con valores nulos son de tipo AGRICULTURE, se ha decidido imputar con la media de las variables de dicha clase:

- Max. Bulding Floor : 3
- Cadastral Quality ID : 4

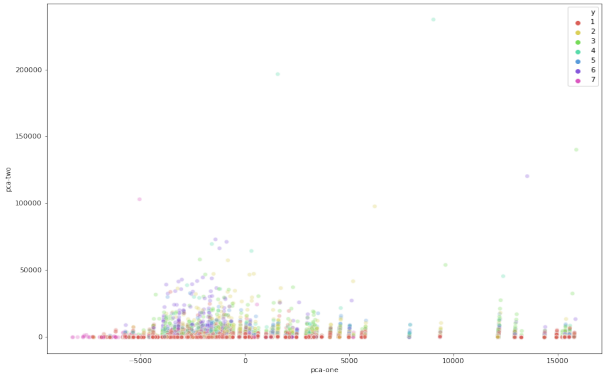
4.2 PCA y TSN-E

Para obtener una intuición sobre como se agrupan las diferentes clases en el espacio dimensional del dataset, se ha realizado un análisis principal de los componentes y se ha aplicado la técnica tsn-e con diferentes perplejidades.

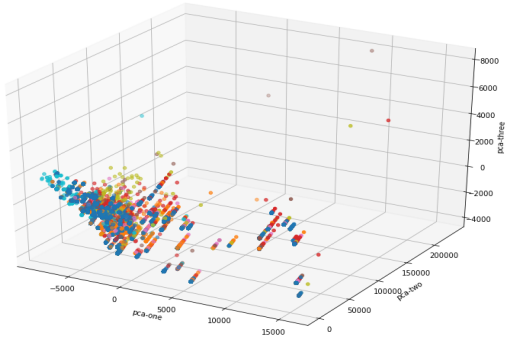
Al aplicar PCA se obtiene los siguientes resultados:

Varianza explicada por componente principal: [0.782 0.109 0.036]

Cuadro 2: PCA 2 Dimensiones



Cuadro 3: PCA 3 Dimensiones



Aunque se pueden observar algunos patrones de agrupamiento, no hay ningún patrón claro que permita distinguir las clases de forma clara, las diferentes clases se superponen mucho entre ellas.

En el Notebook PCA TSN-E[3] se puede ver estos resultados, y los resultados al aplicar diferentes técnicas de balanceamiento a los datos.

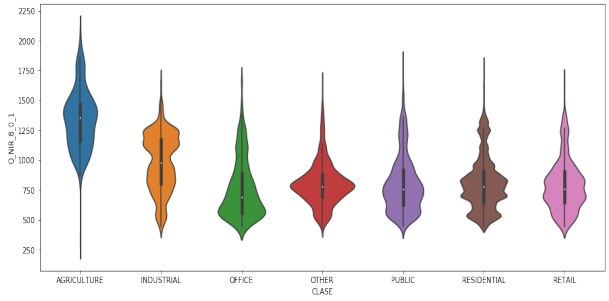
4.3 Distribución variables por clase

Tal y como se muestra en el Notebook ExploratoryVariables[4] las clases se superponen mucho para todas las dimensiones.

No existe ninguna variable que nos permita distinguir la clase perteneciente de forma clara.

Se desarrolla una intución sobre que clases se van a poder distinguir mejor a priori como es el caso de la clase AGRICULTURE y sobre que variables aportan significado o son redundantes.

Cuadro 4: Canal NIR(segundo décil) según clase



5 Feature Engineering

5.1 Variables ID y Coordenadas (X,Y)

Estas variables solo se utilizan para indentificar los registros y en la creación de las variables contexto.

5.2 Reducción de la dimensionalidad

Para cada canal se han eliminado todos los deciles menos el primero, segundo, medio y último.

Esto se ha hecho porque se ha considerado que no aportan ningún significado adicional a las observaciones.

Se ha generado un modelo Random Forest y mediante la importancia de las variables hemos podido confirmar esto. Además mediante el mismo modelo la capacidad predictiva antes y después de eliminar las variables es la misma.

5.3 Variables Categóricas

Las dos variables categóricas que quedan en el dataset se han convertido a numéricas ya que se ha considerado que existe una relación entre las categorias.

5.3.1 Cadastral Quality ID

Según el catastro, la ordenación del código de categoría, de mayor a menor calidad es la siguiente: A, B, C, 1, 2, 3, 4, 5, 6, 7, 8, 9.

Por lo que se ha mapeado cada categoría de forma lineal, siendo el 9 un 1 y A un 12.

5.4 Creación variables contexto

Estás variables se han creado a partir de la intuición de que una finca tiene mas probabilidad de ser de la clase INDUSTRIAL si se encuentra en un polígono industrial y otra tiene poca probabilidad de pertenecer a la clase AGRICULTURE si está en en centro de Madrid.

Aunque las coordenadas X, Y están escaladas y desplazadas y no permiten la ubicar las fincas, si que permiten conocer las fincas vecinas ya que guardan esta relación.

Para cada observación se ha decidido crear una nueva variable por clase (menos RESIDENTIAL), estás variables se llaman contexto_CLASE.

Está variable indica en número (o probabilidad), dependiendo de si la finca se encuentra en Modelar set o Estimar set, de fincas vecinas teniendo en cuenta un número K(4) de vecinos.

Una vez creadas estas nuevas 6 variables se ha demostrado en el notebook Entrega Final[6]que la media de la variable contexto de la clase a la que pertenece es siempre mayor que el resto.

Además con un mismo clasificador se obtienen en diferentes pruebas mejores resultados usando el dataset que contiene estas variables.

6 Modeling

Ya que las clases están muy desbalanceadas y en el dataset de Estimar se presupone que hay menos de la clase mayoritaria RESIDENTIAL, en orden de crear un modelo robusto el balanceamiento de las clases a nivel de datos y/o algoritmo es primordial.

Para ello como métricas objetivo no se tiene solamente en cuanta la del reto, accuracy, sino también el f1 score que tiene en cuenta la precisión y recall. De esta forma y mediante matrices de confusión podemos conocer la performance en las diferentes clases.

Se han probado distintos modelos como SVM, KNN y Redes Neuronales pero siempre se han obtenido mejores resultados con XGBoost.

6.1 Balanceamiento de los datos

En el notebook Diferentes tecnicas de balanceamiento de datos[5] se han estudiado diferentes técnicas de balanceamiento así como sus ventajas, desventajas y performance en orden de poder ser usadas en los futuros modelos.

6.2 Entrega Intermedia

Para está entrega se creo un modelo de Decission Trees mediante la librería XGBoost sobre el dataset que contiene las variables contexto.

Se hizó una búsqueda de los mejores parámetros mediante GridSearch, está búsqueda es muy costosa y por ello se realizó en el cloud.

Se consiguió el segundo puesto en la UPV con una accuracy de 0,6570.

6.3 Entrega Final

Para la presente entrega se ha decidido cambiar el modelo por completo.

El algoritmo seguido sigue las ideas propuestas en este paper Enhancing classification performance of multi-class imbalanced data using the OAA-DB algorithm[12]

El algoritmo es el que sigue:

Las dos capas consisten de 7 clasificadores (uno por clase) siguiendo el modelo uno contra el resto, que predice la probabilidad de que una observación pertenezca a dicha clase.

Cada uno de estos clasificadores es un modelo de predicción binaria creado con XGBoost.

1. En la primera capa los clasificadores se han entrenado sobre el dataset original con las variables contexto.

2. En la segunda capa los clasificadores se han entrenado sobre un dataset sin las variables contexto balanceado con una relación 1:1 mediante la técnica SMOTETomek.

En la primera capa, si solo hay una clase con mas del 15 % de probabilidad, entonces esa es la predicción sino pasamos a la segunda capa.

En la segunda capa, si solo hay una clase con mas del 50 % de probabilidad, entonces esa es la predicción sino la predicción será la clase con más probabilidad según la primera capa.

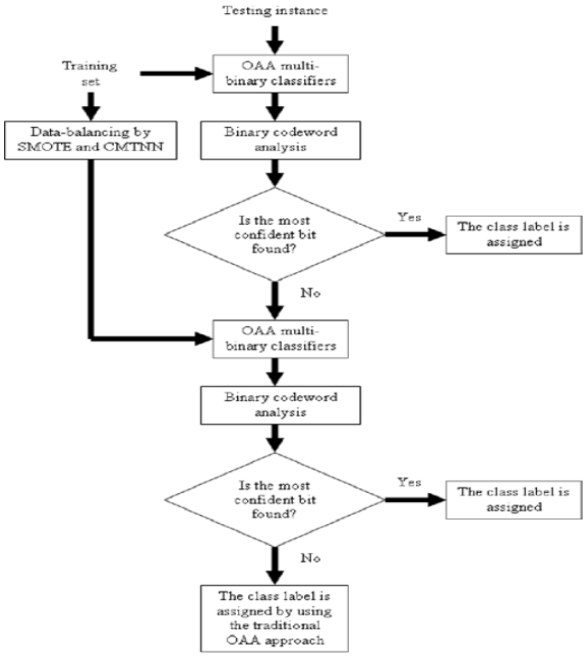
Destacar dos diferencias principales con respecto al algoritmo propuesto en el paper[12], ya que el paper se usan redes neuronales:

- En el algoritmo original la técnica de balanceo de datos es SMOTE para el oversampling y una red neuronal complementaria de veracidad y falsedad CMTNN para el undersampling.

  La CMTNN se encarga de eliminar observaciones difíciles de predecir. Para este fin y por simplicidad yo he optado por usar la técnica de eliminar los Tomek links. Esta técnica elimina observaciones creadas que esten muy próximas en el espacio de todas las dimensiones.

- En el algoritmo original el threshold de la primera capa es del 50 % en vez del 15 % que yo he usado.

Cuadro 5: Flowchart del algoritmo original



Haciendo una partición de train/test de 80/20 se obtienen los siguientes resultados:

Cuadro 6: Resultados

	F1	Accuracy
CLF CAPA 1	0.5127	0.9132
CLF CAPA 2	0.4565	0.7674
MODELO	0.8319	0.9585

7 Trabajo futuro

- Profundizar en feature engineering
- Combinar diferentes modelos mediante ensemble, stacking and bagging
- Optimizar modelos mediante GridSearch

Referencias

Internas

[1] Notebook 1: Data

[2] Notebook 2: Distribución Modelar Estimar

[3] Notebook 3: PCA TSN-E

[4] Notebook 4: Exploratory Variables

[5] Notebook 5: Diferentes tecnicas de balanceamiento de datos

[6] Notebook 6: Entrega Final

Externas

[7] https://sentinel.esa.int/web/sentinel/sentinel-data-access

[8] http://www.sedecatastro.gob.es/

[9] http://www.catastro.minhap.es/documentos/preguntas_frecuentes_formato_CAT.pdf

[10] https://en.wikipedia.org/wiki/Sentinel-2

[11] https://www.agromatica.es/sentinel-2-teledeteccion-agricultura/

[12] https://ieeexplore.ieee.org/document/6252450