Breve resumen del trabajo desarrollado.

Una vez nos enfrentamos al problema, el primer paso fúe el estudio y comprensión del mismo. Realizamos un análisis exploratorio de los datos y calculamos algunas estadísticas sobre ellos. Lo más destacable que encontramos fué que había un gran conflicto entre distintas clases concretas, ya que tenían unas características muy parecidas, siendo de diferentes clases. La primera aproximación que se nos ocurrió fué el uso del paradigma OVO, ya que así podemos enfrentar cada clase frente al resto de clases. La idea principal era obtener un modelo ajustado a cada pareja de clases, sin embargo, debido al tamaño del conjunto de datos y las infraestructuras que disponemos, no era factible la realización de éste ajuste debido al tiempo que podría llegar a consumir. Por ello, decidimos cambiar la estrategia a la utilización del paradigma OVA, que aunque perderíamos la esencia que queríamos del OVO, íbamos a tener un paradigma similar a lo que buscabamos. Una vez obtubimos los modelos, hemos ido realizando diferentes tipos de agregaciones y combinaciones con distintas transformaciones, siendo la mejor la que contaremos en los siguientes apartados.

Resumen análisis exploratorio llevado a cabo y conclusiones.

Lo primero fué calcular las estadísticas básicas de las variables (media, desviación, cuartiles, etc...). Esta información nos servía para tener una idea de cómo se distribuyen las variables. Las variables de color y de inflarrojos no tienen una trasformación inversa, es decir, no podemos reconstruir las imágenes utilizando esa información para así hacer uso de modelos avanzados para extraer información en imágenes. Por lo que en un principio, decidimos no manipular la información que se comparte entre ellas. Las variables relativas a la geometría, parecen ser variables construidas, es decir, no son variables que directamente expresen una característica si no que sobre un conjunto de características sobre el edificio, se ha realizado un algorimo como PCA para crear las variables que aparecen en éste conjunto. Uno de los puntos más claros fué la visualización de los puntos dependiendo de su valor de latitud longitud, en los que pudimos observar, que todas las clases menos 'AGRICULTURE' (y también algunas 'INDUSTRIAL') se pisaban entre ellas y compartían el mismo espacio. Comprobamos el número de instancias por clase, viendo que el problema era un problema muy desbalanceado de la clase 'RESIDENTIAL' frente al resto. Se aplicarán métodos de under_sampling sobre ésta clase para obtener un modelo más competente. Comprobamos que había algunos valores como valores nulos, siendo un procentaje muy pequeño respecto al conjunto, decidimos utilizar simplemente moda para la variable categórica y mediana para las numéricas.

Una vez teníamos todo visualizado y estudiado, decidimos trabajar con el conjunto completo, completando los valores nulos y transformando la variable categórica en numérica ordinal siguiendo el valor numérico de orden que ya nos habían informado con anterioridad.

Como conclusión final, es que las variables tanto de color como geométricas aportan poca información para trabajar con ellas, hicimos una prueba sustituyendo todas las variables de color por el valor medio de cada canal de color y de inflarrojos, con lo que generamos un modelo de igual calidad que con todas las variables, cosa a tener en cuenta para la reducción de dimensionalidad.

Resumen de la manipulación de variables y su argumentación

Como he comentado anteriormente, decidimos realizar una imputación básica de moda para la variable categórica y mediana para las numéricas. Esto simplemente lo hicimos porque había sobre 27 elementos nulos, dentro de 100k instancias, por lo que era mejor aplicar éste mñetodo que alguno más sofisticado.

Aunque hicimos un estudio de correlación y vimos una alta correlación entre variables de color, decidimos dejarlas todas debido a que la posible combinación entre ellas puede ser importante para el modelo.

Por último, la transformación de la variable categórica era necesario para que los algrotimos de aprendizaje puedan trabajar con ella. Debido a que ésta variable tiene una codigificación numérica ordinal, decidimos directamente transformarla a esos valores para así obtener una información más significativa de la misma.

Para compensar la falta de balanceo asignamos a cada clase un peso. La suma de éstos da 1 y se enmarca dentro de las estrategias semisupervisadas, utilizando el conjunto de entrenamiento para estimar la verdadera distribución de los datos de test (éstos últimos se entienden como la verdadera distribución de los datos)

Justificación selección del modelo.

Debido al desbalanceo de clases, se decidió utilizar algorimos o bien basado en árboles, o bien basados en distancias. Por ello, los algoritmos que seleccionamos fueron Random Forest, KNN y XGBoost. Como hemos comentado anteriormente, al tener un alto grado de desbalanceo, decidimos utilizar un algoritmo que en el estado del arte de datos desbalanceados obtiene muy buenos resultados. El algoritmo utilizado fué el EditedNearestNeighbours (ENN). Este algoritmo es un algoritmo de under_sampling, es decir, reduce instancias para balancear la clase. La estrategia que seguimos fué reducir las instancias de la clase mayoritaria en cada modelo de OVA.

Primera aproximacion: OVO:

Una de las estrategias consideradas ha sido la de One-vs-One. Esta estrategia consiste en dividir el problema original de clasificación multiclase de m clases en m*(m-1)/2 subproblemas de clasificación binaria, de forma que en cada uno de ellos se tratará de clasificar cada pareja de clases, generando de esta forma un modelo para cada pareja. Una vez realizado esto, se deberán de agregar los resultados de los diferentes clasificadores para obtener una predicción final.

Esta estrategia nos presenta una serie de ventajas teóricas que pensamos que coinciden con las necesidades del problema en cuestión. En primer lugar, es una de las técnicas recurrentes en la literatura, y se ha demostrado que puede obtener buenos resultados en los problemas de clasificación no balanceada[1]. Al tratarse nuestro problema de una problema altamente desbalanceado, pensamos que es una estratégia que nos pueda hacer llegar a un buen resultado.

En segundo lugar, la descomposición del problema en susbproblemas nos ofrece la ventaja de preprocesar y tratar los datos de forma diferente para cada pareja de clases. De esta forma podremos intentar encontrar el preprocesamiento óptimo para cada par. En este sentido, se ha llevado a cabo una selección y creación de características, prubas con diferentes algoritmos de clasificación y ajuste de parámetros para cada par de clases.

La selección de características llevada a cabo en este caso consiste en una selección mediante un algoritmo greedy, el cual tratará de dar con un subconjunto de características que optimice la separación de ambas clases. Por otra parte, se han tratado de crear nuevas características a partir de aquellas que parecen más prometedoras, haciendo uso también de la técnica greedy, y aplicando transformaciones polinómicas y logarítmicas.

Por último, en relación a esta técnica, tendremos que decantarnos por una forma de agregar los resultados arrojados por cada clasificador y generar la predicción final. Para ello, se ha hecho uso de algunas de las posibles agregaciones expuestas en [2], entre las cuales podemos ver estrategias de voto simple, voto ponderado en función de la probabilidad de pertenencia a cada clase, y otras agregaciones más complejas.

A pesar de todas las ventajas teóricas que esta técnica nos ofrece, los resultados obtenidos no consiguen mejorar de forma sustancial los resultados obtenidos con otras técnicas más simples en lo que a número de modelos necesarios se refiere, como One-vs-All, por lo tanto esta no ha sido nuestra solución final. Pensamos que uno de los problemas que empeora la calidad del resultado es la elección de una agregación que no se comporta bien en este problema. Este será uno de los principales puntos a mejorar para la proxima fase de la competición.

Segunda aproximación: OVA: Lo que hicimos fué ajustar un modelo para cada clase enfrentada a la unión del resto de clases. El proceso es el siguiente:

- Se transforman y se escalan los datos.
- Se realiza under_sampling para intentar balancear el conjunto algo más de lo que estaba.
- Se ajusta un modelo cualquiera de los anteriores (el que mejor resultado obtenga en validación).

Una vez se hace este proceso de forma iterativa tenemos las probabilidades de pertenecer a cada una de las clases, calculadas con el mejor modelo individual obtenido

en una batería de pruebas. Lo que hacemos ahora es simplemente seleccionar la clase que mayor probabilidad ha obtenido. Se han probado distintos métodos como seleccionar los k vecinos más cercanos y eliminar la clase que no se encuentre entre esos k vecinos y también con un Naive Bayes para que determine el orden de selección en vez de seleccionar la mayor probabilidad. Ninguno de éstos métodos ha mejorado y por lo tanto el que se utilizó fué el de seleccionar el máximo.

Aproximación final: OVA + Calibración:

Una vez obtenido el OVA, para calibrar la confianza de los modelos aplicamos un desplazamiento constante por columna del OVA obtenido mediante diferential evolution asignando los pesos a clases antes descritos.

[1] Zhang, Zhongliang & Krawczyk, Bartosz & García, Salvador & Rosales-Pérez, Alejandro & Herrera, Francisco. (2016). Empowering One-vs-One Decomposition with Ensemble Learning for Multi-Class Imbalanced Data. Knowledge-Based Systems. 106. 10.1016/j.knosys.2016.05.048.

[2] M. Galar, A. Fernández, E. Barrenechea, H. Bustince and F. Herrera, An Overview of Ensemble Methods for Binary Classifiers in Multi-class Problems: Experimental Study on One-vs-One and One-vs-All Schemes. Pattern Recognition 44:8 (2011) 1761-1776, doi: 10.1016/j.patcog.2011.01.017.