**Evaluación de técnicas de balanceo por medio de Mapas Autoorganizativos**

# INTRODUCCIÓN

El objetivo de este trabajo es evaluar diferentes técnicas de balanceo de datos aplicadas sobre datasets de clasificación binaria, los cuales poseen una distribución de clases muy desequilibrada, esto con frecuencia provoca que los algoritmos de Machine Learning no puedan extraer información de forma correcta y tiendan a producir errores. Para evaluar el correcto funcionamiento de estas técnicas a la hora de balancear un conjunto de datos, vamos a utilizar Mapas Autoorganizativos de Kohonen, entrenándolos previamente con los datos originales y clasificando posteriormente los datos tras aplicar todo este conjunto de técnicas. A partir de aquí se calcularan diversos tipos de errores sobre los mapas para validar el buen desempeño de estas técnicas.

# DATOS

Tenemos un total de 4 conjuntos de datos desbalanceados extraídos del libro *¨Imbalanced Classification with Python¨* sobre los que vamos a trabajar. En el apartado de proyectos de dicho libro, podremos encontrar una descripción de cada uno de los conjuntos de datos junto a la descripción de las variables que lo forman con las que entrenaremos los mapas. Vamos a usar los siguientes datasets:

* Supervivencia de pacientes de cáncer de mama.
* Detección de derrames de petróleo en el mar.
* Asignación de créditos bancarios.
* Clasificación de fonemas.

# TÉCNICAS APLICADAS

Para solucionar el problema de desbalanceo de las categorías que se quieren predecir se van a aplicar las siguientes técnicas de balanceo, habiendo tanto métodos de oversampling como de undersampling, las cuales serán combinadas y aplicadas sobre los conjuntos de datos. Todas las técnicas se encuentran descritas y referenciadas en el libro *¨Imbalanced Classification for Python¨* a excepción de la técnica de *Kmeans SMOTE* (Last, Douzas, & Bacao, 2017)*.* En la carpeta de ´SOTA´ se encuentran estas referencias y algunas más relacionadas con el proyecto.

* Como técnicas de oversampling usaremos:
  + SMOTE
  + ADASYN
  + Borderline SMOTE
  + SVM SMOTE
  + Kmeans SMOTE
* Como técnicas de undersampling usaremos:
  + Tomek Links
  + Edited Nearest Neighbors
  + Condensed Nearest Neighbors
  + Neighbourhood Cleaning Rule
  + One Side Selection

A la hora de normalizar los datos se ha aplicado la técnica de MinMax, debido a que con la normalización Euclidea los datos resultantes para algunos de los datasets son muy pequeños y provocan errores al realizar los cálculos del error topológico y el error de cuantificación de los mapas.

# FLUJO DE TRABAJO PROPUESTO

Para la realización de la comparativa y las pruebas se ha establecido el flujo de trabajo que podemos ver en la *Ilustración 1*:

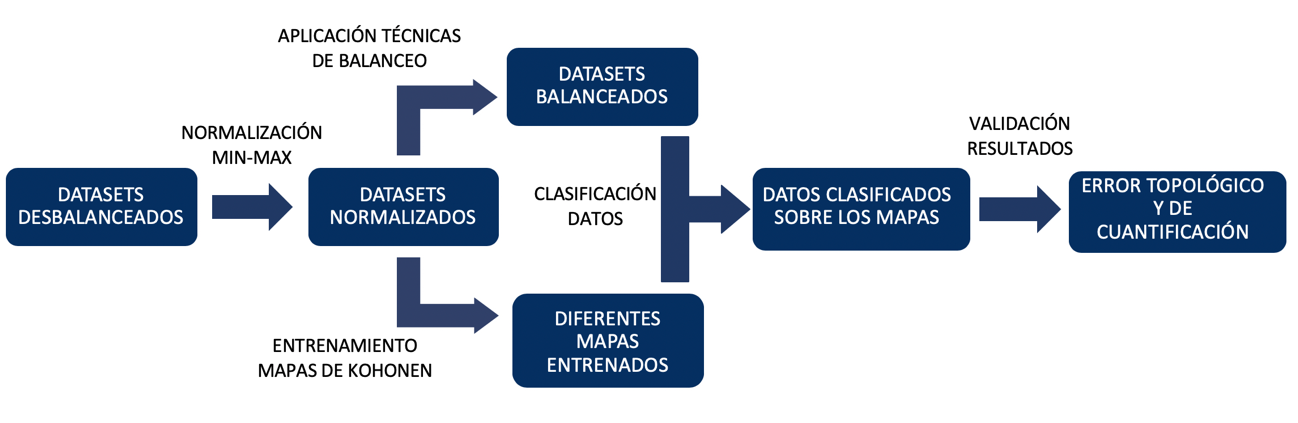


Ilustración 1: Flujo de trabajo establecido para los experimentos

Primero vamos a normalizar los conjuntos de datos originales de cara a representarlos todos bajo la misma escala. A continuación, por un lado entrenaremos los mapas con los datos originales, probando las diferentes combinaciones de parámetros que podemos ver en la *Tabla 1* y por el otro, aplicaremos todas las combinaciones de técnicas sobre los conjuntos de datos. A continuación clasificaremos los datos balanceados sobre los mapas entrenados con los datos originales y por último, calcularemos los errores de cuantificación de topológico de los mapas.

|  |  |
| --- | --- |
| **Parámetros** | |
| **Lado del mapa** | 5 – 25\* |
| **Periodo** | 500, 1000, 2500, 5000, 7500, 10000 |
| **Learning Rate** | 0.01, 0.05, 0.1, 0.2, 0.3 |

Tabla 1: Parámetros para el entrenamiento de los mapas

\*Los valores de lado de mapa de cada uno de los experimentos pueden variar en función del lado de mapa óptimo calculado.

# RESULTADOS

Para los resultados, se ha seleccionado de forma fija el lado de mapa óptimo aproximado para cada de los conjuntos de datos, calculado con la siguiente fórmula:

Siendo:

* = Número de neuronas aproximado del mapa de Kohonen.
* = Número de instancias totales con los que vamos a entrenar al mapa.

Para la realización de las pruebas tenemos un notebook por cada uno de los datasets que tenemos. En el mismo se implementa el flujo de trabajo descrito previamente en la *Ilustración 1* y por cada combinación de las técnicas de balanceo se genera un archivo .csv con los errores de cuantificación y topológico y las activaciones del mapa para cada combinación de posibles parámetros de la *Tabla 1*.

En la parte final de los notebooks también está implementada la posibilidad de entrenar un único mapa con los mejores parámetros que se hayan obtenido previamente y estudiar el mismo.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | | **Dataset – cáncer mama** | | | | |
|  | | | | **Lado Mapa** | **Periodo** | | **Learning Rate** | **Error de clasificación (media y desviación típica al clasificar los datos sintéticos)** | **Error Cuantificación** | **Error topológico** |
| **SMOTE** | **Tomek Links** | | | 9 | 5000 | | 0.30 |  |  |  |
| **Edited Nearest Neighbors** | | | 9 | 7500 | | 0.30 |  |  |  |
| **Condensed Nearest Neighbors** | | | 9 | 7500 | | 0.30 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 9 | 5000 | | 0.30 |  |  |  |
| **One Side Selection** | | | 9 | 2500 | | 0.30 |  |  |  |
| **ADASYN** | **Tomek Links** | | | 9 | 7500 | | 0.10 |  |  |  |
| **Edited Nearest Neighbors** | | | 9 | 10000 | | 0.30 |  |  |  |
| **Condensed Nearest Neighbors** | | | 9 | 7500 | | 0.20 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 9 | 7500 | | 0.20 |  |  |  |
| **One Side Selection** | | | 9 | 5000 | | 0.30 |  |  |  |
| **Borderline SMOTE** | **Tomek Links** | | | 9 | 7500 | | 0.20 |  |  |  |
| **Edited Nearest Neighbors** | | | 9 | 7500 | | 0.30 |  |  |  |
| **Condensed Nearest Neighbors** | | | 9 | 7500 | | 0.30 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 9 | 7500 | | 0.20 |  |  |  |
| **One Side Selection** | | | 9 | 10000 | | 0.30 |  |  |  |
| **SVM SMOTE** | **Tomek Links** | | | 9 | 5000 | | 0.30 |  |  |  |
| **Edited Nearest Neighbors** | | | 9 | 7500 | | 0.20 |  |  |  |
| **Condensed Nearest Neighbors** | | | 9 | 7500 | | 0.20 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 9 | 7500 | | 0.30 |  |  |  |
| **One Side Selection** | | | 9 | 7500 | | 0.30 |  |  |  |
| **Kmeans SMOTE** | **Tomek Links** | | | 9 | 7500 | | 0.20 |  |  |  |
| **Edited Nearest Neighbors** | | | 9 | 10000 | | 0.20 |  |  |  |
| **Condensed Nearest Neighbors** | | | 9 | 7500 | | 0.30 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 9 | 5000 | | 0.20 |  |  |  |
| **One Side Selection** | | | 9 | 7500 | | 0.20 |  |  |  |

Tabla 2: Resultados de la combinación de técnicas para el dataset de cáncer de mama

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | | **Dataset – derrames petróleo** | | | | |
|  | | | | **Lado Mapa** | **Periodo** | | **Learning Rate** | **Error de clasificación (media y desviación típica al clasificar los datos sintéticos)** | **Error Cuantificación** | **Error topológico** |
| **SMOTE** | **Tomek Links** | | | 12 | 7500 | | 0.20 |  |  |  |
| **Edited Nearest Neighbors** | | | 12 | 1000 | | 0.30 |  |  |  |
| **Condensed Nearest Neighbors** | | | 12 | 5000 | | 0.30 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 12 | 7500 | | 0.30 |  |  |  |
| **One Side Selection** | | | 12 | 500 | | 0.20 |  |  |  |
| **ADASYN** | **Tomek Links** | | | 12 | 1000 | | 0.30 |  |  |  |
| **Edited Nearest Neighbors** | | | 12 | 500 | | 0.20 |  |  |  |
| **Condensed Nearest Neighbors** | | | 12 | 1000 | | 0.30 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 12 | 1000 | | 0.20 |  |  |  |
| **One Side Selection** | | | 12 | 500 | | 0.20 |  |  |  |
| **Borderline SMOTE** | **Tomek Links** | | | 12 | 500 | | 0.30 |  |  |  |
| **Edited Nearest Neighbors** | | | 12 | 1000 | | 0.30 |  |  |  |
| **Condensed Nearest Neighbors** | | | 12 | 2500 | | 0.30 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 12 | 2500 | | 0.30 |  |  |  |
| **One Side Selection** | | | 12 | 1000 | | 0.20 |  |  |  |
| **SVM SMOTE** | **Tomek Links** | | | 12 | 2500 | | 0.20 |  |  |  |
| **Edited Nearest Neighbors** | | | 12 | 2500 | | 0.20 |  |  |  |
| **Condensed Nearest Neighbors** | | | 12 | 2500 | | 0.30 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 12 | 1000 | | 0.30 |  |  |  |
| **One Side Selection** | | | 12 | 2500 | | 0.30 |  |  |  |
| **Kmeans SMOTE** | **Tomek Links** | | | 12 | 5000 | | 0.30 |  |  |  |
| **Edited Nearest Neighbors** | | | 12 | 5000 | | 0.30 |  |  |  |
| **Condensed Nearest Neighbors** | | | 12 | 2500 | | 0.20 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 12 | 2500 | | 0.20 |  |  |  |
| **One Side Selection** | | | 12 | 1000 | | 0.30 |  |  |  |

Tabla 3: Resultados de la combinación de técnicas para el dataset de derrames de petróleo

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | | **Dataset – créditos bancarios** | | | | |
|  | | | | **Lado Mapa** | **Periodo** | | **Learning Rate** | **Error de clasificación (media y desviación típica al clasificar los datos sintéticos)** | **Error Cuantificación** | **Error topológico** |
| **SMOTE** | **Tomek Links** | | | 13 | 2500 | | 0.20 |  |  |  |
| **Edited Nearest Neighbors** | | | 13 | 5000 | | 0.10 |  |  |  |
| **Condensed Nearest Neighbors** | | | 13 | 5000 | | 0.10 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 13 | 2500 | | 0.20 |  |  |  |
| **One Side Selection** | | | 13 | 7500 | | 0.10 |  |  |  |
| **ADASYN** | **Tomek Links** | | | 13 | 5000 | | 0.30 |  |  |  |
| **Edited Nearest Neighbors** | | | 13 | 7500 | | 0.30 |  |  |  |
| **Condensed Nearest Neighbors** | | | 13 | 7500 | | 0.10 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 13 | 5000 | | 0.20 |  |  |  |
| **One Side Selection** | | | 13 | 7500 | | 0.10 |  |  |  |
| **Borderline SMOTE** | **Tomek Links** | | | 13 | 5000 | | 0.30 |  |  |  |
| **Edited Nearest Neighbors** | | | 13 | 5000 | | 0.20 |  |  |  |
| **Condensed Nearest Neighbors** | | | 13 | 2500 | | 0.30 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 13 | 2500 | | 0.30 |  |  |  |
| **One Side Selection** | | | 13 | 7500 | | 0.20 |  |  |  |
| **SVM SMOTE** | **Tomek Links** | | | 13 | 2500 | | 0.30 |  |  |  |
| **Edited Nearest Neighbors** | | | 13 | 2500 | | 0.20 |  |  |  |
| **Condensed Nearest Neighbors** | | | 13 | 7500 | | 0.10 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 13 | 5000 | | 0.30 |  |  |  |
| **One Side Selection** | | | 13 | 5000 | | 0.10 |  |  |  |
| **Kmeans SMOTE** | **Tomek Links** | | | 13 | 5000 | | 0.30 |  |  |  |
| **Edited Nearest Neighbors** | | | 13 | 7500 | | 0.20 |  |  |  |
| **Condensed Nearest Neighbors** | | | 13 | 7500 | | 0.20 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 13 | 7500 | | 0.30 |  |  |  |
| **One Side Selection** | | | 13 | 2500 | | 0.30 |  |  |  |

Tabla 4: Resultados de la combinación de técnicas para el dataset de créditos bancarios

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | | **Dataset – fonemas** | | | | |
|  | | | | **Lado Mapa** | **Periodo** | | **Learning Rate** | **Error de clasificación (media y desviación típica al clasificar los datos sintéticos)** | **Error Cuantificación** | **Error topológico** |
| **SMOTE** | **Tomek Links** | | | 19 | 7500 | | 0.10 |  | 0.077 |  |
| **Edited Nearest Neighbors** | | | 19 | 7500 | | 0.30 |  |  |  |
| **Condensed Nearest Neighbors** | | | 19 | 7500 | | 0.20 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 19 | 7500 | | 0.10 |  |  |  |
| **One Side Selection** | | | 19 | 7500 | | 0.30 |  |  |  |
| **ADASYN** | **Tomek Links** | | | 19 | 10000 | | 0.20 |  |  |  |
| **Edited Nearest Neighbors** | | | 19 | 7500 | | 0.20 |  |  |  |
| **Condensed Nearest Neighbors** | | | 19 | 7500 | | 0.30 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 19 | 1000 | | 0.10 |  |  |  |
| **One Side Selection** | | | 19 | 7500 | | 0.30 |  |  |  |
| **Borderline SMOTE** | **Tomek Links** | | | 19 | 10000 | | 0.20 |  |  |  |
| **Edited Nearest Neighbors** | | | 19 | 10000 | | 0.20 |  |  |  |
| **Condensed Nearest Neighbors** | | | 19 | 10000 | | 0.20 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 19 | 7500 | | 0.30 |  |  |  |
| **One Side Selection** | | | 19 | 7500 | | 0.20 |  |  |  |
| **SVM SMOTE** | **Tomek Links** | | | 19 | 7500 | | 0.20 |  |  |  |
| **Edited Nearest Neighbors** | | | 19 | 7500 | | 0.30 |  |  |  |
| **Condensed Nearest Neighbors** | | | 19 | 7500 | | 0.30 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 19 | 10000 | | 0.20 |  |  |  |
| **One Side Selection** | | | 19 | 10000 | | 0.30 |  |  |  |
| **Kmeans SMOTE** | **Tomek Links** | | | 19 | 7500 | | 0.20 |  |  |  |
| **Edited Nearest Neighbors** | | | 19 | 7500 | | 0.30 |  |  |  |
| **Condensed Nearest Neighbors** | | | 19 | 10000 | | 0.30 |  |  |  |
| **Neighbourhood Cleaning Rule** | | | 19 | 10000 | | 0.20 |  |  |  |
| **One Side Selection** | | | 19 | 10000 | | 0.30 |  |  |  |

Tabla 5: Resultados de la combinación de técnicas para el dataset de fonemas