Imbalanced Data and Label Uncertaintly

Conjuntos de datos desbalanceados y etiquetado con incertidumbre

Contenido

[Imbalanced Data 1](#_Toc168311096)

[1. Introducción 1](#_Toc168311097)

[2. Naturaleza del problema 1](#_Toc168311098)

[Label Uncertaintly 3](#_Toc168311099)

[Referencias 4](#_Toc168311100)

## Imbalanced Data

### Introducción

A medida que avanza la ciencia y la tecnología, la capacidad de recolectar datos se ha incrementado significativamente. Sin embargo, la presencia de datos desbalanceados, es decir, que las etiquetas de los datos no son proporcionales en el dataset, puede provocar que las técnicas de *machine learning* (ML) estándar vean afectado su rendimiento y su capacidad de predicción.

### Naturaleza del problema

Cualquier conjunto de datos con una proporción desigual entre las clases se puede considerar desbalanceado, pero comúnmente se entiende por conjunto de datos desbalanceado cuando existe una diferencia significativa en la proporción de las etiquetas.

Suponiendo un conjunto de datos de mamografías (He & Garcia, 2009) con 10,923 “Negativos” (clase mayoritaria) y 260 “Positivos” (clase minoritaria) en relación con la presencia de cáncer, se desearía obtener un clasificador que determine la etiqueta sin errar en la predicción, con un 100% de acierto. En la realidad, el clasificador tenderá a clasificar con la etiqueta de la clase mayoritaria a prácticamente todo el conjunto de datos, obteniendo un grado de acierto cercano al 100%. Sin embargo, el grado de acierto en las observaciones que pertenecen a la clase contraria sería cercano a 0. En el ejemplo, 234 pacientes de cáncer de mama serían clasificados como sanos, lo que provocaría un error catastrófico para los pacientes y médicos, ya que el objetivo en medicina es detectar la clase minoritaria para poder ofrecer el tratamiento. Además, el tipo de error también es importante, ya que el coste de determinar erróneamente como cancerosa a una persona sana es mucho menor que el de diagnosticar como sana a una persona que realmente tiene cáncer. El clasificador ideal proporcionaría una alta precisión para la clase minoritaria sin afectar a la precisión de la clase mayoritaria.

Por esta razón, métricas convencionales como el *general* *accuracy* o el *error rate*, no proporcionan información adecuada para modelos entrenados con conjuntos de datos desbalanceados. Por tanto, es necesario utilizar métricas de evaluación más informativas como receiver *operating characteristic curve* (ROC), *precision-recall* y curvas de coste.

La presencia de datos desbalanceados puede deberse a la naturaleza inherente de los datos en el espacio (desbalanceo intrínseco) o a factores externos, como el tiempo o el almacenamiento de los datos, que no están directamente relacionados con la naturaleza del espacio de los datos (desbalanceo extrínseco).

Además, existen dos tipos de desbalanceo: el relativo y la rareza absoluta. El desbalanceo por rareza absoluta se debe a contar con muy pocas observaciones, tan pocas que parecen datos atípicos. El desbalanceo relativo entre las clases indica que existe un mayor número de unidades de una clase frente a la otra, aunque, que esto ocurra, no quiere decir que no existan datos suficientes de la clase minoritaria. El desbalanceo relativo es muy común en el mundo real y el centro de muchos esfuerzos de investigación.

Estudios demuestran que la efectividad de un modelo entrenado con un conjunto de datos relativamente desequilibrado no es de por si contraproducente, ya que el principal factor es la complejidad del conjunto de datos, la perturbación del desequilibrio (superposición, falta de datos representativos, pequeñas disyuntivas, etc.) el que puede afectar a la clasificación, amplificado si existe además desbalanceo relativo. En la *Figura 1*, (a) el desbalanceo relativo no tiene superposición y solo tiene un concepto por cada clase. Por otro lado, (b) tiene múltiples conceptos y se observa superposición. Además, el subconcepto C puede pasar desapercibido por algunos clasificadores al poderse considerar casos raros.

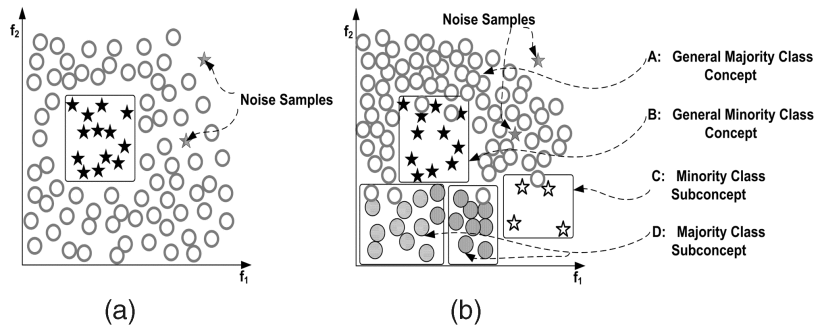


Figura 1. (a) Dataset con desbalanceo relativo. (b) Alta complejidad en el desbalanceo relativo. Fuente: Learning from Imbalanced Data

El desequilibrio por *casos raros* es representativo de dominios donde los ejemplos de clases minoritarias son muy limitados, siendo el concepto objetivo raro y dificultando el aprendizaje independientemente del desequilibrio entre clases. Además, el concepto minoritario puede contener un subconjunto de instancias limitadas, lo que implica diferentes grados de dificultad en la clasificación. Este hecho causa una nueva forma de desequilibrio llamado *desequilibrio dentro de la clase*, que se centra en la distribución de datos representativos para subconceptos dentro de una clase.

La existencia de un *desequilibrio dentro de la clase* está relacionada con el problema de *pequeñas disyuntivas* que empeoran el rendimiento del clasificador ya que este intentará aprender un concepto mediante la creación de múltiples reglas disyuntivas que describan el concepto principal. Para el caso de conceptos homogéneos, el clasificador creará grandes disyuntivas y para el caso de conjuntos heterogéneos, las pequeñas disyuntivas surgen como resultado directo de subconceptos subrepresentados. Además, el ruido puede influir en las disyuntivas de la clase minoritaria y la validez de los *clusters*, siendo cuestión importante determinar si estos representan un subconjunto real o si es debido al ruido.

Un último problema para los clasificadores es la combinación de datos desbalanceados

## Label Uncertaintly

## Referencias

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