TRABAJO PRACTICO N° 5- PARTE 1

9) En el caso de que un dataset se encuentre desbalanceado existen muchos métodos para solucionar este inconveniente, pero los más usados son el método de sobremuestreo y l de submuestreo que se detallaran a continuación

Métodos de sobremuestreo: Consisten en la creación de nuevas muestras de clase minoritaria. Los métodos más usados son 2: a) crear las muestras al azar las muestras minoritarias y a) SMOTE esta técnica básicamente generar más muestras de la clase minoritaria a partir de las que ya tenemos.

Método de submuestreo: se basa en quitar muestra de la clase mayoritaria para intentar equilibrar el número de muestras en cada clase. La técnica más simple es el submuestreo aleatorio que implica la eliminación aleatoria de los ejemplos de la clase mayoritaria hasta balancear la base de datos

TRABAJO PRACTICO N° 5- PARTE 2

|  |  |  |  |
| --- | --- | --- | --- |
| DATASET | ERROR | PRESICION | SENSIBIILIDAD |
| BALANCEADO 160 DATOS POR CLASE- METODO SVM | **0.207650273224044** | **0.8087432** | **0.9925926** |
| BALANCEADO 320 DATOS POR CLASE- METODO SVM | **0.119891008174387** | **0.880109** | **0.9815498** |
| DESBALANCEADO- METODO SVM | **0.0880626223091976** | **0.9119374** | **0** |
| BALANCEADO 160 DATOS POR CLASE- METODO RF | **0.3125** | **0.6875** | **0.6875** |
| BALANCEADO 320 DATOS POR CLASE- METODO RF | **0.760416666666667** | **0.7604167** | **0.7272727** |
| DESBALANCEADO- METODO RF | **0.102678571428571** | **0.8973214** | **0.5714286** |

3) Métricas de precisión

Es una métrica de evaluación utilizada en problemas de clasificación que mide la proporción de casos positivos identificados correctamente por el modelo sobre el total de casos identificados como positivos.

La fórmula para calcular la precisión es la siguiente:

Precisión = Verdaderos positivos / (Verdaderos positivos + Falsos positivos)

Donde:

* Verdaderos positivos (TP, True Positives): número de casos positivos que el modelo ha clasificado correctamente.
* Falsos positivos (FP, False Positives): número de casos negativos que el modelo ha clasificado incorrectamente como positivos.

**La precisión es importante cuando el costo de un falso positivo es alto**, es decir, cuando es importante minimizar los casos en que se clasifica incorrectamente un caso negativo como positivo. Por ejemplo, en un problema de detección de [**spam**](https://es.wikipedia.org/wiki/Spam)en el correo electrónico, es importante minimizar los casos en que un correo electrónico legítimo es clasificado como spam.

Sin embargo, la precisión no considera los casos que el modelo no ha identificado correctamente como positivos, es decir, los falsos negativos (FN, False Negativos). Por lo tanto, es importante utilizar otras métricas de evaluación en conjunto con la precisión para tener una visión más completa del rendimiento del modelo en la tarea de clasificación.

Sensibilidad

La sensibilidad (recall en inglés) es una métrica de evaluación utilizada en problemas de clasificación que mide la proporción de casos positivos identificados correctamente por el modelo sobre el total de casos positivos presentes en los datos.

La fórmula para calcular la sensibilidad es la siguiente:

Sensibilidad = Verdaderos positivos / (Verdaderos positivos + Falsos negativos)

Donde:

* Verdaderos positivos (TP, True Positives): número de casos positivos que el modelo ha clasificado correctamente.
* Falsos negativos (FN, False Negativos): número de casos positivos que el modelo ha clasificado incorrectamente como negativos.

La sensibilidad es importante cuando el costo de un falso negativo es alto, es decir, cuando es importante identificar correctamente todos los casos positivos presentes en los datos. Por ejemplo, en un problema de [diagnóstico médico](https://es.wikipedia.org/wiki/Diagn%C3%B3stico_m%C3%A9dico), es importante identificar correctamente todos los casos de una enfermedad para poder proporcionar el tratamiento adecuado.

La sensibilidad no considera los casos que el modelo ha identificado incorrectamente como positivos, es decir, los falsos positivos (FP, False Positives). Por lo tanto, es importante utilizar otras métricas de evaluación en conjunto con la sensibilidad para tener una visión más completa del rendimiento del modelo en la tarea de clasificación.

4) Conclusiones:

* A mayor conjunto de datos en el dataset mejora la predicción
* A mayor cantidad de datos en el dataset disminuye el error
* El método SVM cuando se aplicó al dataset desbalanceado el valor de la sensibilidad da cero quiere decir que clasificó mal toda una clase
* El método SVM funciona mejor con un dataset balanceado ya que cuando se aplica a un dataset desbalanceado el error es mayor que cuando lo hace en un dataset balanceado
* El método RF funciona mejor con un dataset desbalanceado ya que cuando se aplica a un dataset balanceado el error es mayor que cuando lo hace en un dataset desbalanceado

Link de GitHub : https://github.com/danielmartincampos/FIA-TP5