SMOTE

Quando abbiamo a che fare con dataset sbilanciati bisogna stare attenti perché i metodi tradizionali di classificazione assumono che il dataset sia bilanciato fra le sue classi.
Per bilanciare un dataset, si possono usare diverse metodologie:

1) OVERSAMPLING
2) UNDERSAMPLING
3) MUOVERE LA THRESHOLD: così che le tuple della classe minoritaria siano più facili da classificare.
4) ENSEMBLE METHODS

Synthetic Minority Over-Sampling Technique.

Il principio è quello di creare sinteticamente delle tuple per la classe minoritaria. Questi nuovi punti sintetici vengono creati sui segmenti che congiungono il punto corrente e i suoi K-NN della stessa classe.
Dipendentemente da quanti punti sintetici si vogliono creare, si sceglie un sottoinsieme dei K-NN da cui generare i dati. (Se abbiamo K=5 e vogliamo che i punti diventino il 200% in più allora basta selezionare solo i due vicini più vicini per ogni punto, così da creare due punti sintetici per ogni punto reale).

PASSI

1) Calcolare la differenza fra due punti della stessa classe (differenza fra feature vectors).

2) Moltiplicare questa differenza per un numero random, così da avere un punto nuovo lungo il segmento che unisce i due punti reali.

3) Aggiungerlo al dataset

NOTA

Se bisogna fare un ribilanciamento, bisogna farlo solo nel Training Set!