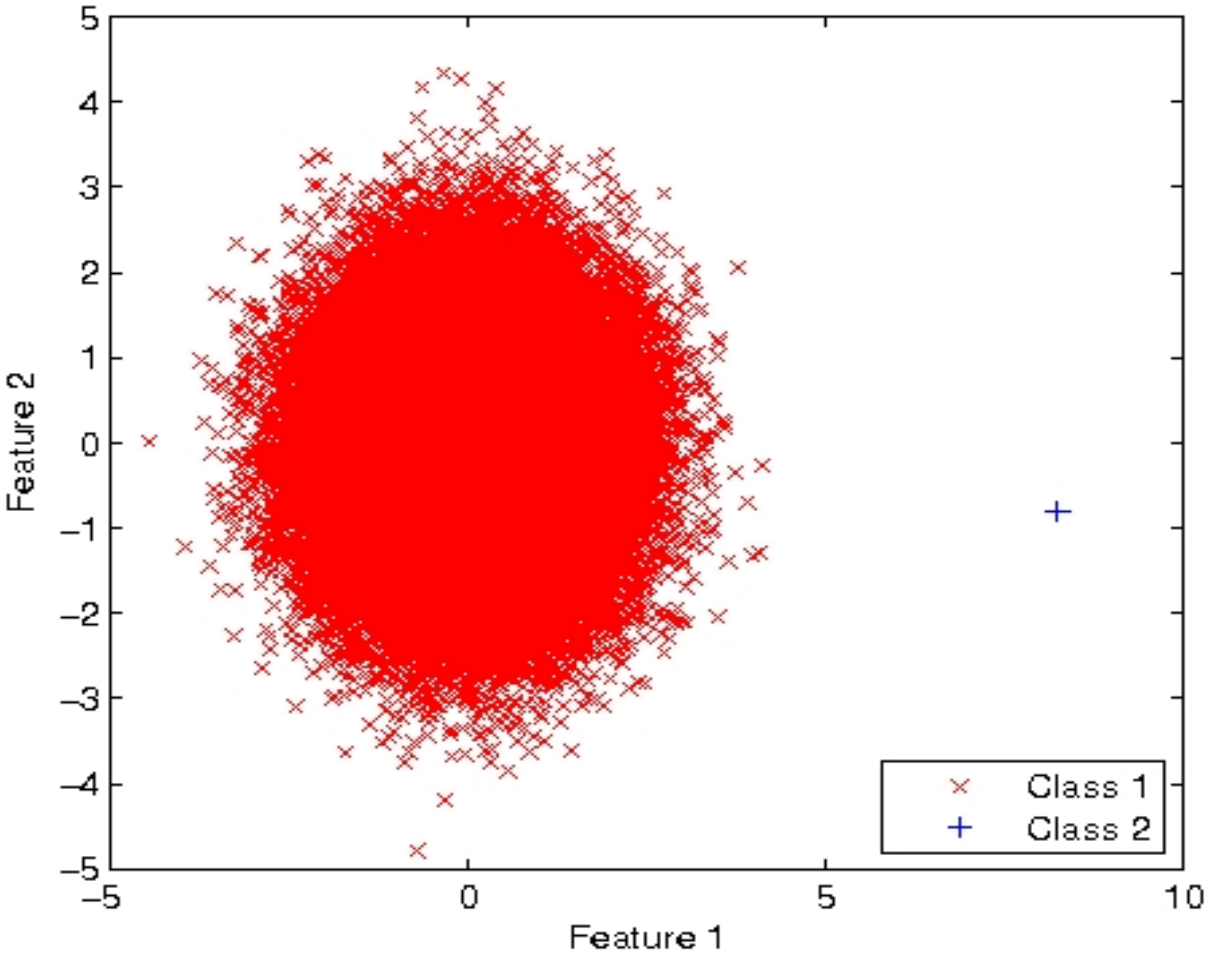
SMOTE

SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

PROBLEM DANYCH NIEZBALANSOWANYCH

Problem danych niezbalansowanych pojawia się wtedy, gdy liczność jednej klasy (klasy dominującej, przyjmuje się że jest to klasa negatywna) jest istotnie wyższa niż liczność drugiej klasy (klasy zdominowanej, pozytywnej). Istotą problemu niezbalansowania jest fakt, że zastosowanie klasycznych mechanizmów uczenia na niezrównoważonym zbiorze danych może prowadzić do faworyzowania przez wyuczony klasyfikator klasy dominującej kosztem klasy zdominowanej.

PROBLEM DANYCH NIEZBALANSOWANYCH



PROBLEM DANYCH NIEZBALANSOWANYCH

Do rozwiązania problemu niezbalansowania stosuje metody przetwarzania danych polegające na próbkowaniu z klasy zdominowanej (oversampling) bądź też eliminacji obserwacji z klasy dominującej (undersampling). Do najpopularniejszych metod zalicza się:

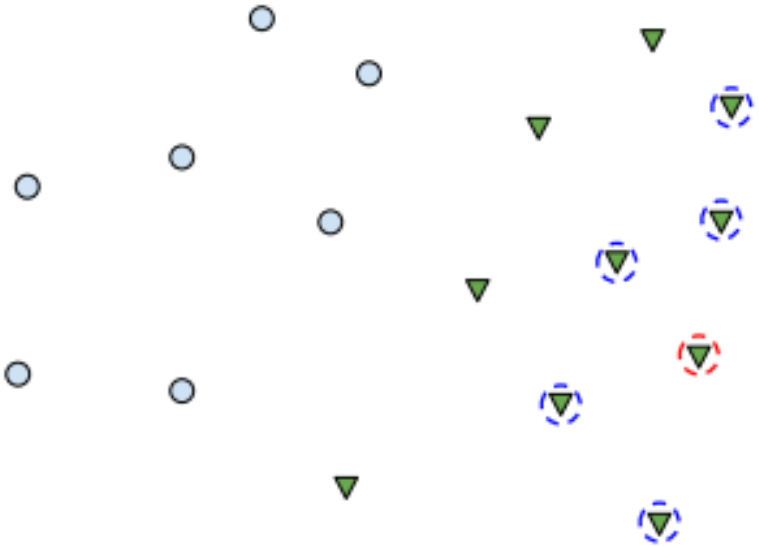
- **Algorytm SMOTE**
- **Algorytm Selekcji Jednostronnej**

ALGORYTM SELEKCJI JEDNOSTRONNEJ

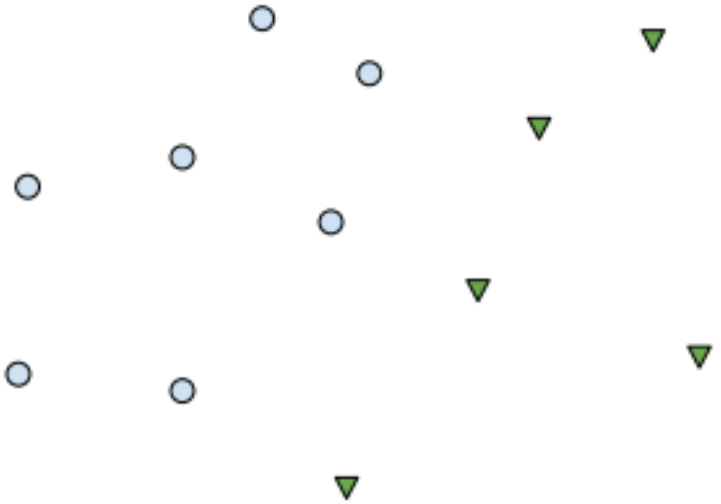
Tę metodę definiuje się w następujących krokach:

- **Usuń ze zbioru uczącego wszystkie obserwacje z klasy dominującej.**
- **Wylosuj jedną obserwację *x* z klasy dominującej i dodaj do zbioru uczącego.**
- **Dla każdej z pozostałych obserwacji z klasy dominującej x_n sprawdź, czy x_n jest bliżej x, czy też bliżej jest dowolna obserwacja z klasy zdominowanej.**
- **Jeżeli bliżej x_n znajduje się obserwacja z klasy zdominowanej (a NIE x należący do dominującej) to dodaj x_n do zbioru.**

ALGORYTM SELEKCJI JEDNOSTRONNEJ



(a) Obserwacja wybrana w wyniku losowania (kolor czerwony) i obserwacje wybrane do eliminacji (kolor niebieski).



(b) Dane po wykonaniu selekcji jednostronnej.

ALGORYTM SMOTE

Algorytm SMOTE jest metodą generowania syntetycznych próbek z klasy zdominowanej. Syntetyczne próbki umieszczane są na odcinku łączącym najbliżej położone siebie obiekty z jednej i drugiej klasy.

Sposób działania algorytmu prezentuje pseudokod.

ALGORYTM SMOTE

```
Wejście
  T - liczba przykładów z klasy zdominowanej
  N - procentowa liczebność zbioru SMOTEd względem zbioru oryginalnego
  k - liczba najbliższych sasiadów
Wyjście:
  (N/100)*T - elementowy zbiór przykładów SMOTEd z klasy zdominowanej

SMOTE(T, N, k)
begin
  /* Jeśli N jest mniejsze od 100, losowy zbiór przykładów zostanie zrównoważony */
  if N < 100 then
    Losuj przykłady z klasy T zdominowanej
    T = (N/100) * T
    N = 100
  endif
  N = (int)(N/100) /* N jest liczbą naturalną będącą wielokrotnością 100 */
  k = Liczba najbliższych sąsiadów
  numattrs = Liczba atrybutów
  Sample[ ][ ]: Macierz oryginalnych obiektów klasy zdominowanej
  newindex: zapamiętuje liczbę wygenerowanych syntetycznych przykładów, inizjalizowany jako 0
  Synthetic[ ][ ]: Macierz przykładów syntetycznych
  /* Odnajduje k njbliższych sąsiadów dla każdego obiektu klasy zdominowanej */
  for i <- 1 to T
    Odnajduje k najbliższych sasiadów dla i, zapisuje do tablicy nnarray
    Populate(N, i, nnarray)
  endfor
end
```

ALGORYTM SMOTE

```
Populate(N, i, nnarray) /* Funkcja do generowania przykładów syntetycznych. */
begin
  while N != 0
    Wybierz losową liczbę ze zbioru od 1 do k i przypisz są do zmiennej nn. Ten krok wybiera
jednego z najbliższych sąsiadów i.
    for attr <- 1 to numattrs
      Oblicz: dif = Sample[nnarray[nn]][attr] - Sample[i][attr]
      Oblicz: gap = losowa liczba z przediału od 0 do 1
      Synthetic[newindex][attr] = Sample[i][attr] + gap * dif
    endfor
    newindex++
    N = N - 1
  endwhile
  return
end
```

ALGORYTM SMOTE

Rozważmy przykład (6,4) i załóżmy, że (4,3) jest jego najbliższym sąsiadem.

(6,4) jest przykładem, dla którego określono k najbliższych sąsiadów.

(4,3) jest jednym z k najbliższych sąsiadów

f1_1 = 6, f2_1 = 4, f2_1 – f1_1 = -2

f1_2 = 4, f2_2 = 3, f2_2 – f1_2 = -1

Nowe przykłady zostaną wygenerowane według:

(f1',f2') = (6,4) + rand(0-1)*(-2,-1)

gdzie rand(0-1) oznacza losową liczbę z przedziału 0-1

ALGORYTM SMOTE + UNDERSAMPLING

Często metodę SMOTE poprzedza się dodatkowo operacją under-samplingu. Polega to na tym, że z klasy dominującej usuwa się przykłady dopóki liczebność klasy zdominowanej nie stanie się pewnym określonym ułamkiem klasy dominującej.