SMOTE

Wilfried Segretier
wsegreti@univ-ag.fr

M1 Miage 2018-2019
Université des Antilles

28 novembre 2018

Sommaire

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction

Classification et déséquilibre entre classes

SMOTE

1 Introduction

2 Classification et déséquilibre entre classes

3 SMOTE

Introduction

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction
Classification et déséquilibre entre classes
SMOTE

Classification

- Nécessité d'identifier des classes
- Données d'apprentissage : création de modèles
- Données représentatives et équilibrées
- Problème : monde réel déséquilibre entre classes
- Exemples : Crue/Non Crue, Fraude/Non Fraude, Malade/Sain, ...
- Coût des erreurs différents

Exemple : Iris

Wilfried Segretier
wsegreti@univ-ag.fr

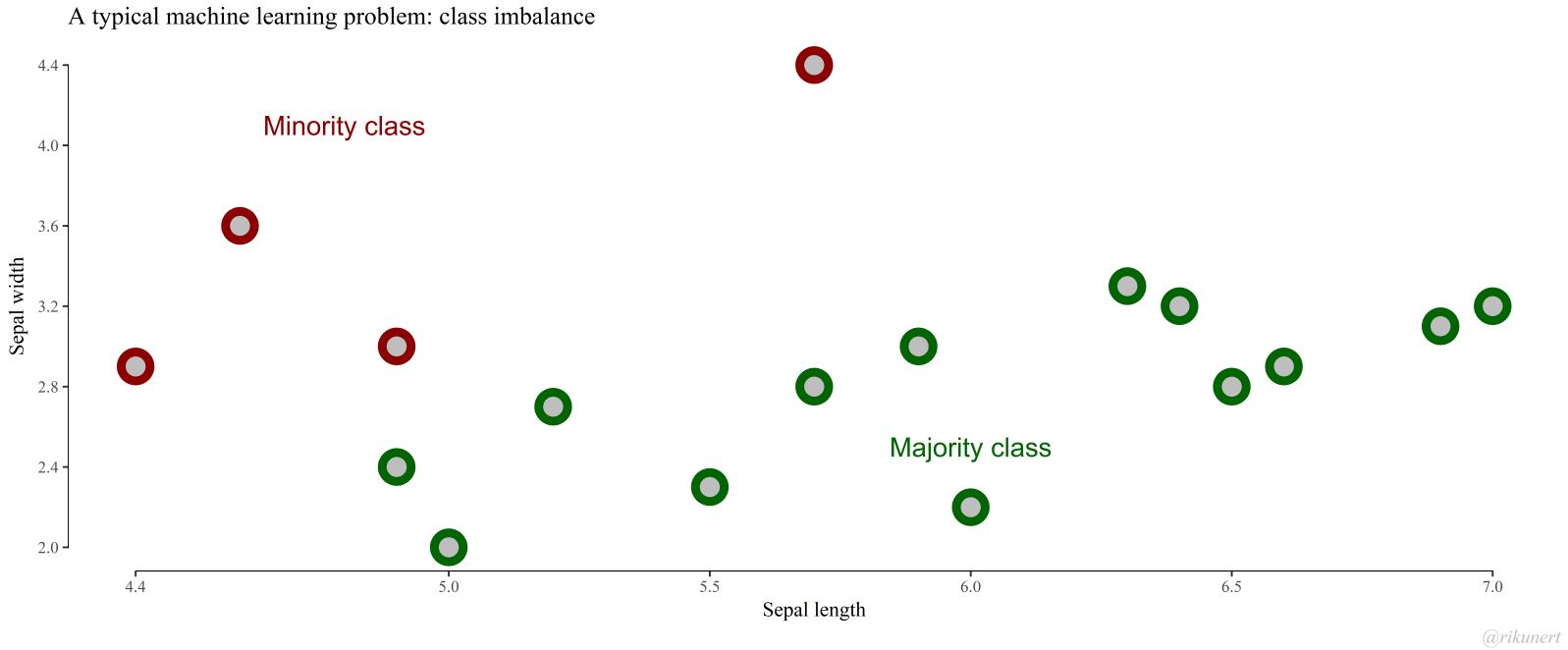
Introduction
Classification et déséquilibre entre classes
SMOTE

```
> dat
   Sepal.Length Sepal.Width    Species
2           4.9         3.0     setosa
9           4.4         2.9     setosa
16          5.7         4.4     setosa
23          4.6         3.6     setosa
51          7.0         3.2 versicolor
52          6.4         3.2 versicolor
53          6.9         3.1 versicolor
54          5.5         2.3 versicolor
55          6.5         2.8 versicolor
56          5.7         2.8 versicolor
57          6.3         3.3 versicolor
58          4.9         2.4 versicolor
59          6.6         2.9 versicolor
60          5.2         2.7 versicolor
61          5.0         2.0 versicolor
62          5.9         3.0 versicolor
63          6.0         2.2 versicolor
```

Exemple : Iris

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction
Classification et déséquilibre entre classes
SMOTE

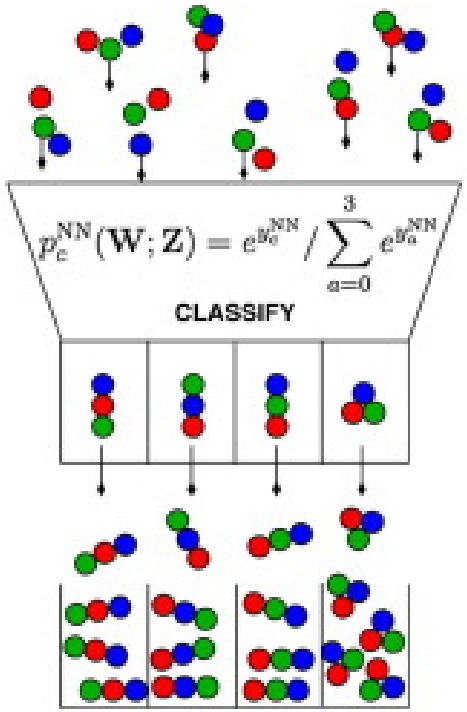


Classification

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction
Classification et déséquilibre entre classes
SMOTE

- extraction de modèles prédictifs à partir des données
- discrimination de classes
- apprentissage, test, validation
- évaluation des performances
- $\neq$ types : arbres, réseaux neurones, modèles bayesiens, règles d'association, ...



Classification : Matrice de confusion

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction
Classification et déséquilibre entre classes
SMOTE

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Precision = (T P + T N)/(T P + F P + T N + F N)
%T P = T P/(T P +F N)
%F P = F P/(T N+F P)

Stratégie contre le déséquilibre entre classes

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction
Classification et déséquilibre entre classes
SMOTE

- Sous-échantillonage de la classe majoritaire
- Sur-échantillonage de la classe minoritaire avec replacement
- Mise en oeuvre d'une fonction de coût augmentant la pondération des erreurs sur la classe minoritaire
- Génération de nouveau exemple de la classe minoritaire : SMOTE

Fonctions de coût : exemples

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction
Classification et déséquilibre entre classes
SMOTE

$$F_\beta = (1+\beta^2) \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$$

$$IV = \sum (P(S_i|A) - P(S_i|B) * log\left(\frac{P(S_i|A)}{P(S_i|B)}\right)$$

$$IV = (TN - FN) * log\left(\frac{TN}{FN}\right) + (FP - TP) * log\left(\frac{FP}{TP}\right)$$

SMOTE

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction
Classification et déséquilibre entre classes
SMOTE

Synthetic Minority Over-sampling TEchnique

- Chawla et al., 2002
- Méthode permettant de générer de nouveaux exemples à partir des exemples minoritaires
- Paramétrable, simple
- Utilisée avec succès pour de nombreux cas d'étude
- Peut etre comninée avec d'autres approches (under-sampling)

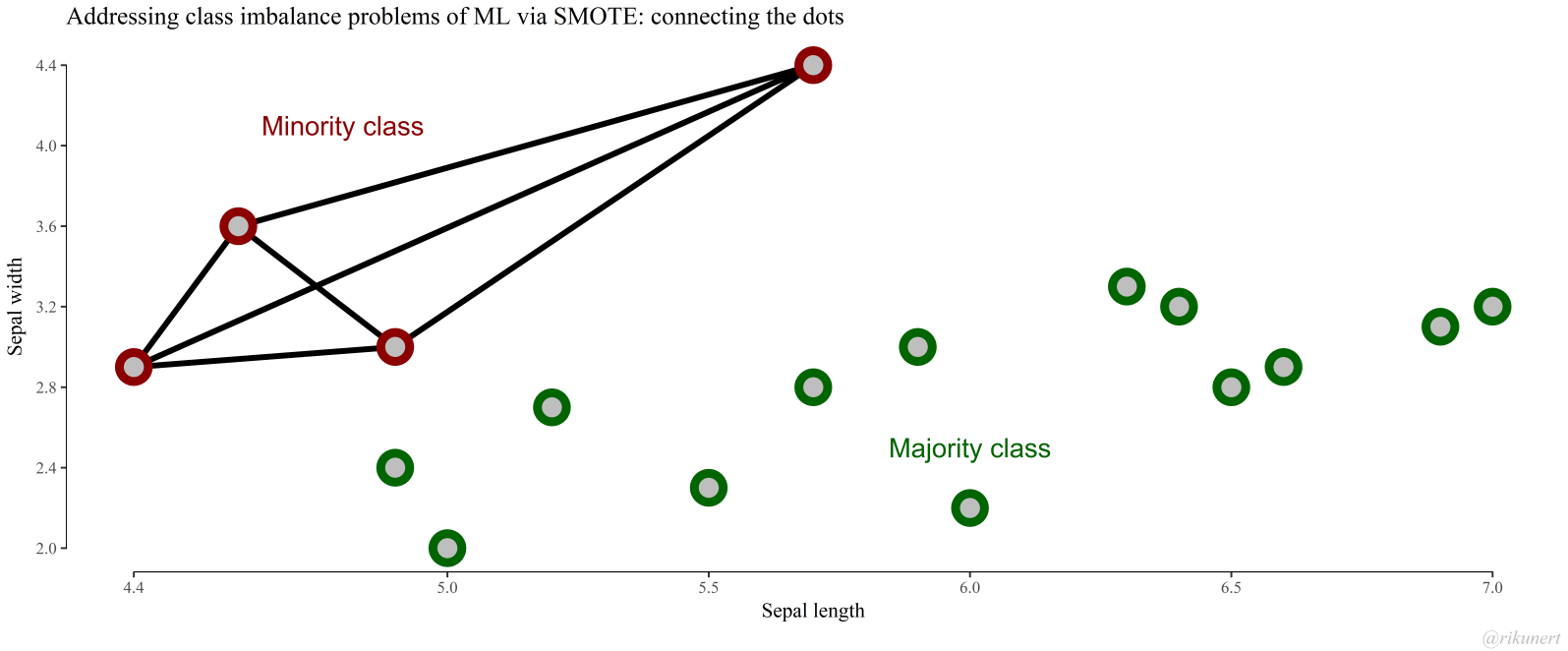
Exemple : Iris

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction

Classification et déséquilibre entre classes

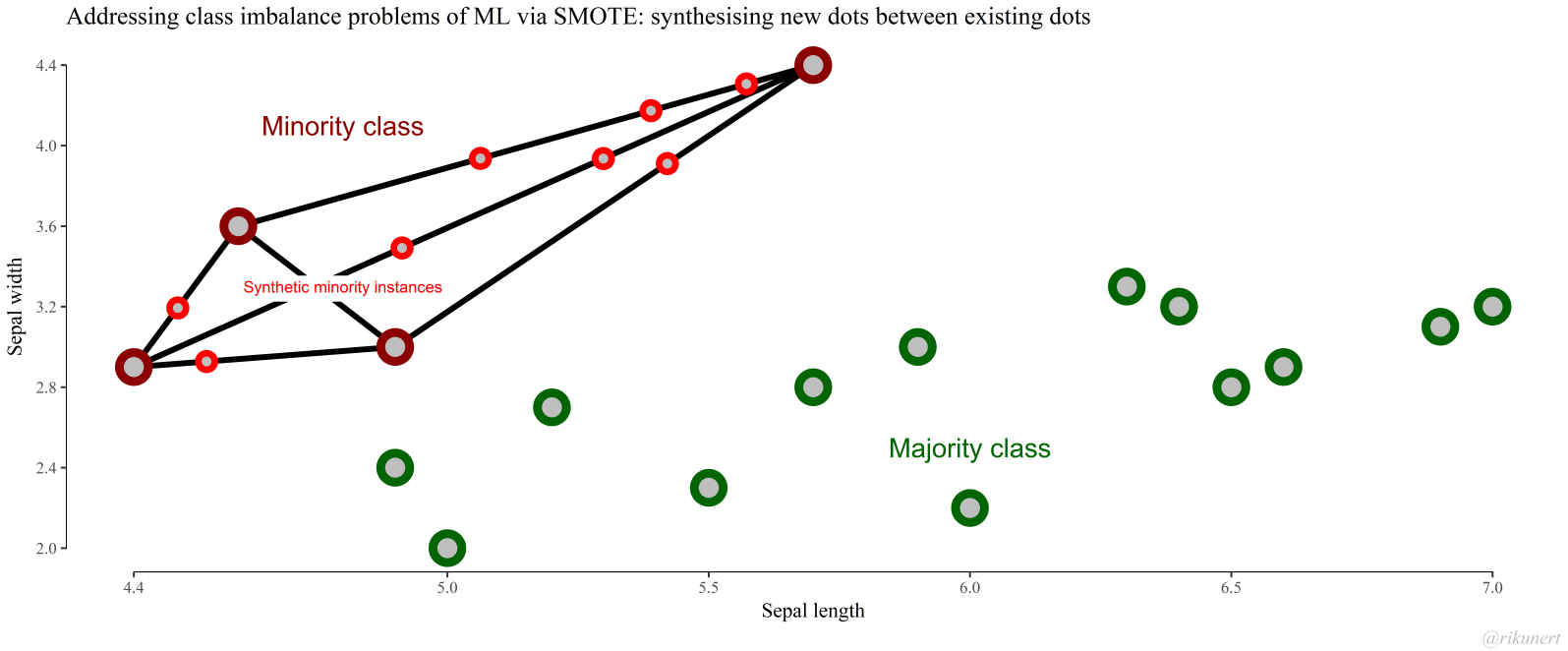
SMOTE



Exemple : Iris

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction
Classification et déséquilibre entre classes
SMOTE



Smote : Paramétrage

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction
Classification et déséquilibre entre classes
SMOTE

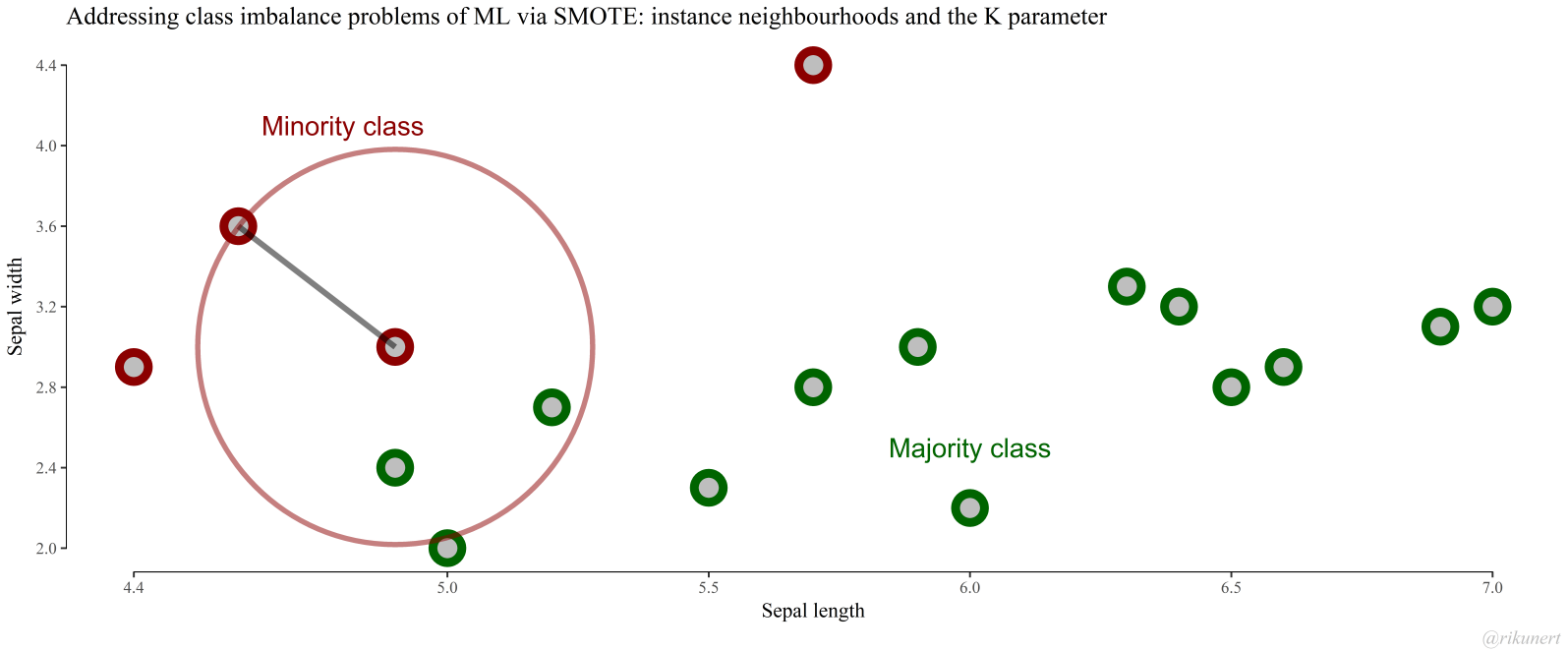
2 paramètres

1. K
   - nombre de plus proches voisins considérés lors de la création de nouveaux exemples
   - 1, 2, 3, ...
   - 0 : recherche valeur optimale
2. dup_size
   - nombre d'itérations pour chaque exemple de la classe minoritaire
   - 1 : 4 nouveaux exemples
   - 200 : 800 nouveaux exemples

Exemple : Iris

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction
Classification et déséquilibre entre classes
SMOTE



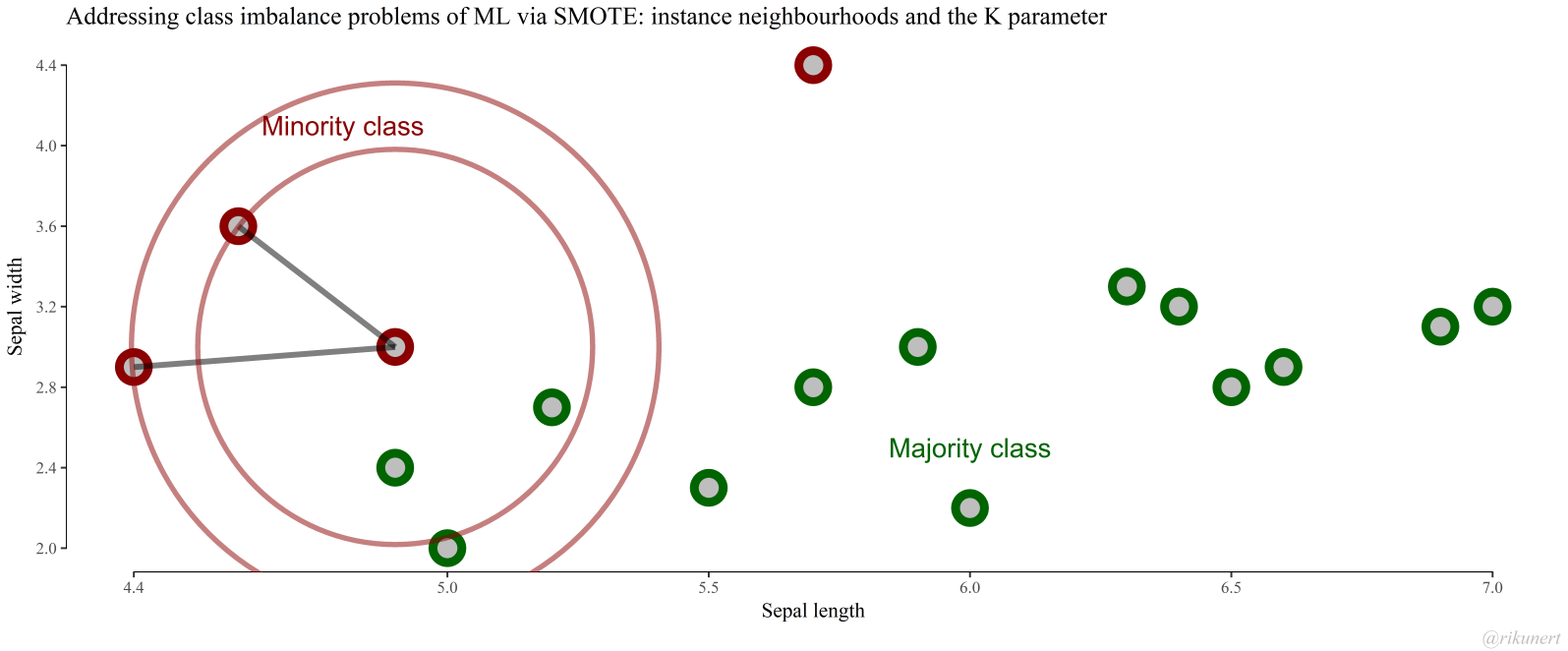
Exemple : Iris

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction

Classification et déséquilibre entre classes

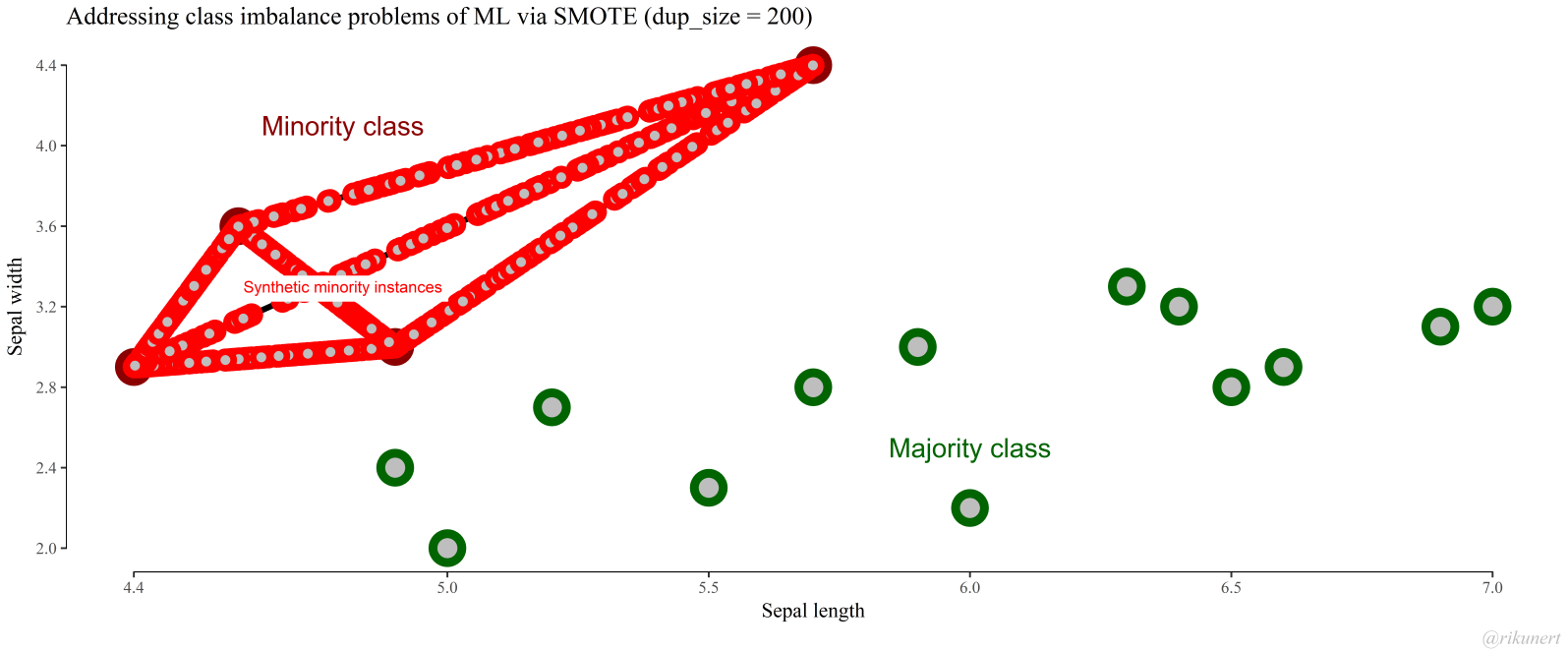
SMOTE



Exemple : Iris

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction
Classification et déséquilibre entre classes
SMOTE



SMOTE : Algorithme 1/2

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction
Classification et déséquilibre entre classes
SMOTE

Algorithm *SMOTE*(T, N, k)
Input: Number of minority class samples T; Amount of SMOTE $N\%$; Number of nearest neighbors k
Output: $(N/100)$ * T synthetic minority class samples

1. (∗ *If N is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTEd.* ∗)
2. **if** $N < 100$
3. **then** Randomize the T minority class samples
4. $T = (N/100) * T$
5. $N = 100$
6. **endif**
7. $N = (int)(N/100)$ (∗ *The amount of SMOTE is assumed to be in integral multiples of 100.* ∗)
8. k = Number of nearest neighbors
9. $numattrs$ = Number of attributes
10. $Sample[\,][\,]$: array for original minority class samples
11. $newindex$: keeps a count of number of synthetic samples generated, initialized to 0
12. $Synthetic[\,][\,]$: array for synthetic samples
(∗ *Compute k nearest neighbors for each minority class sample only.* ∗)
13. **for** $i \leftarrow 1$ **to** T
14. Compute k nearest neighbors for i, and save the indices in the $nnarray$
15. Populate(N, i, $nnarray$)
16. **endfor**

SMOTE : Algorithme 2/2

Populate(N, i, *nnarray*) (∗ *Function to generate the synthetic samples.* ∗)
while $N \neq 0$
    Choose a random number between 1 and k, call it nn. This step chooses one of the k nearest neighbors of i.
    for $attr \leftarrow 1$ **to** $numattrs$
        Compute: $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$
        Compute: gap = random number between 0 and 1
        $Synthetic[newindex][attr] = Sample[i][attr] + gap * dif$
    endfor
    $newindex{+}{+}$
    $N = N - 1$
endwhile
return (∗ *End of Populate.* ∗)
End of Pseudo-Code.

Template

Wilfried Segretier
wsegreti@univ-ag.fr

Introduction

Classification et déséquilibre entre classes

SMOTE