Supervised classification of satellite image time series from noisy labeled data.

Charlotte PELLETIER[a]
charlotte.pelletier@cesbio.cnes.fr

Silvia VALERO[a]
silvia.valero@cesbio.cnes.fr

Nicolas CHAMPION[b]
nicolas.champion@ign.fr

Jordi INGLADA[a]
jordi.inglada@cesbio.cnes.fr

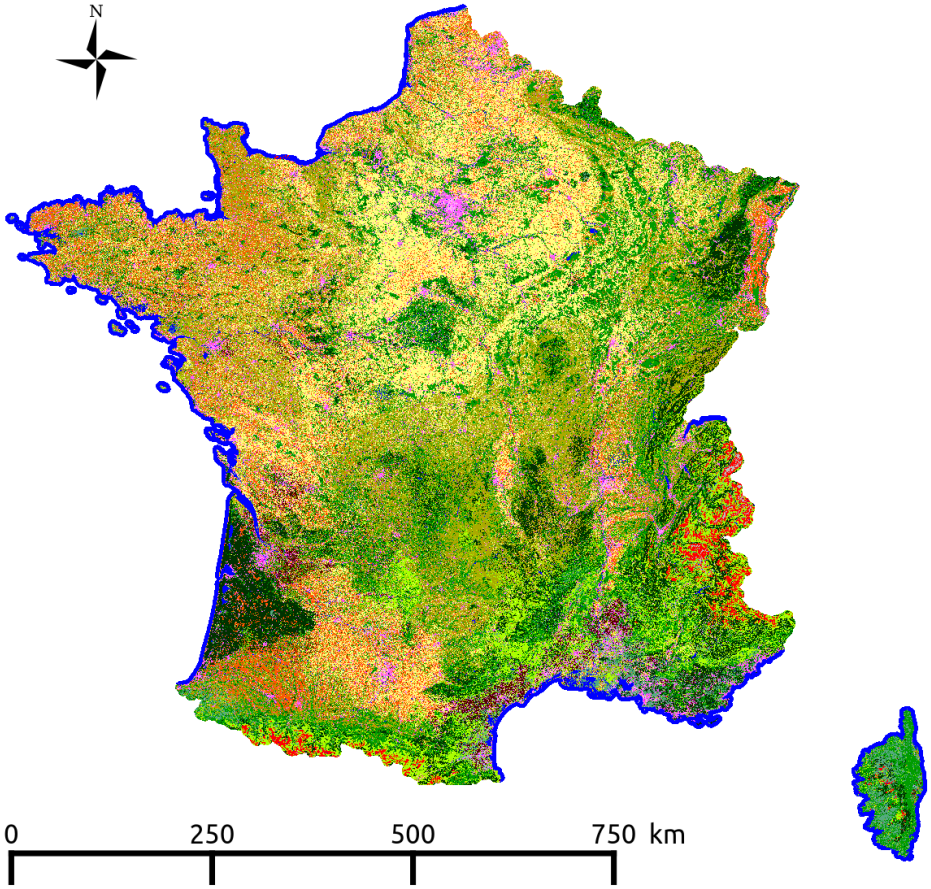
Gérard DEDIEU[a]
gerard.dedieu@cesbio.cnes.fr

(a) CESBIO - UMR 5126, Université Paul Sabatier, CNES, CNRS, IRD - 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, FRANCE
(b) IGN Espace - MATIS - Université Paris-Est, 6 avenue de l'Europe, 31521 Ramonville Cedex, FRANCE

01/12/2017

Context

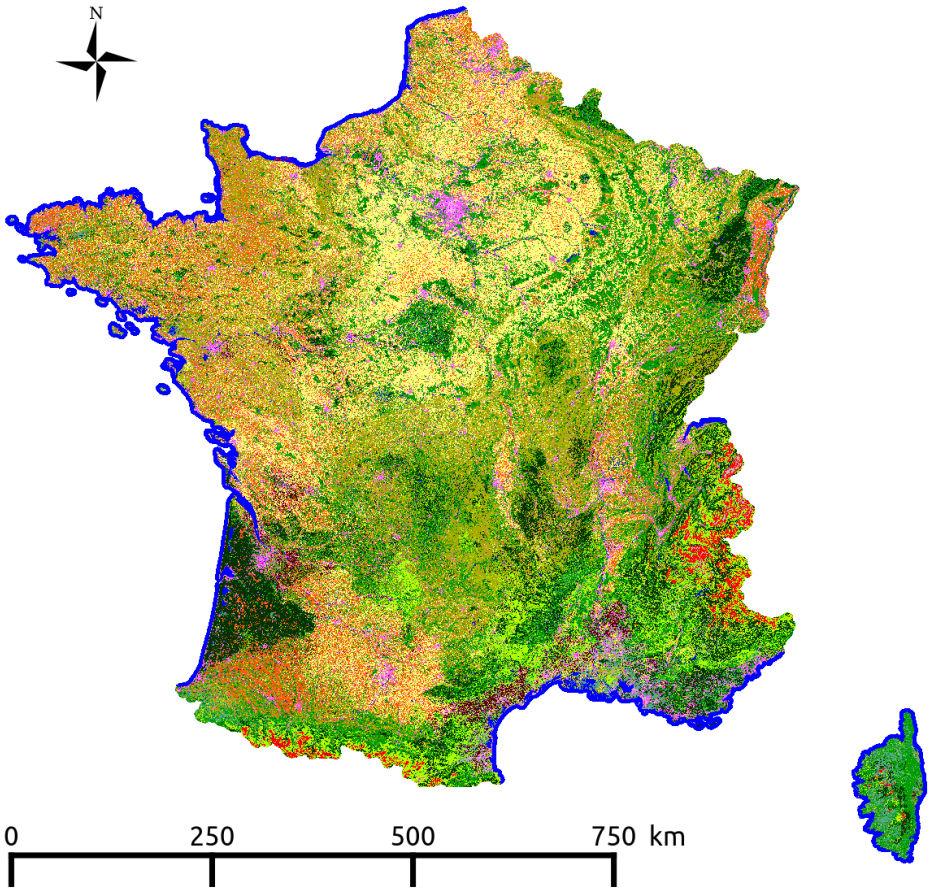
Land cover mapping over large areas covering 15/20 classes:



`http://osr-cesbio.ups-tlse.fr/%7Eoso/`
(Inglada *et al.*, 2017)

Context

Land cover mapping over large areas covering 15/20 classes:



http://osr-cesbio.ups-tlse.fr/%7Eoso/
(Inglada *et al.*, 2017)

Goal

- automatic production
- at national scale
- with a frequent update

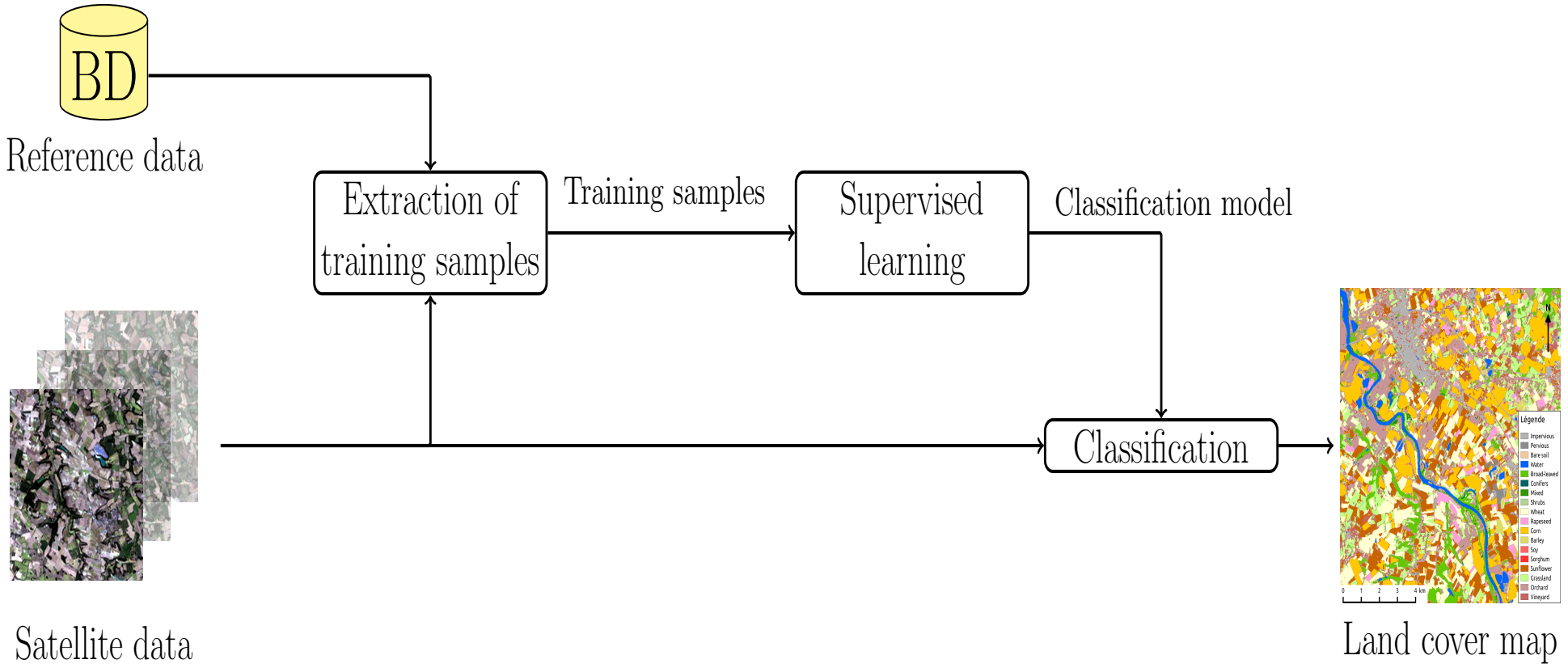
Context

- new satellite image time series at high resolutions

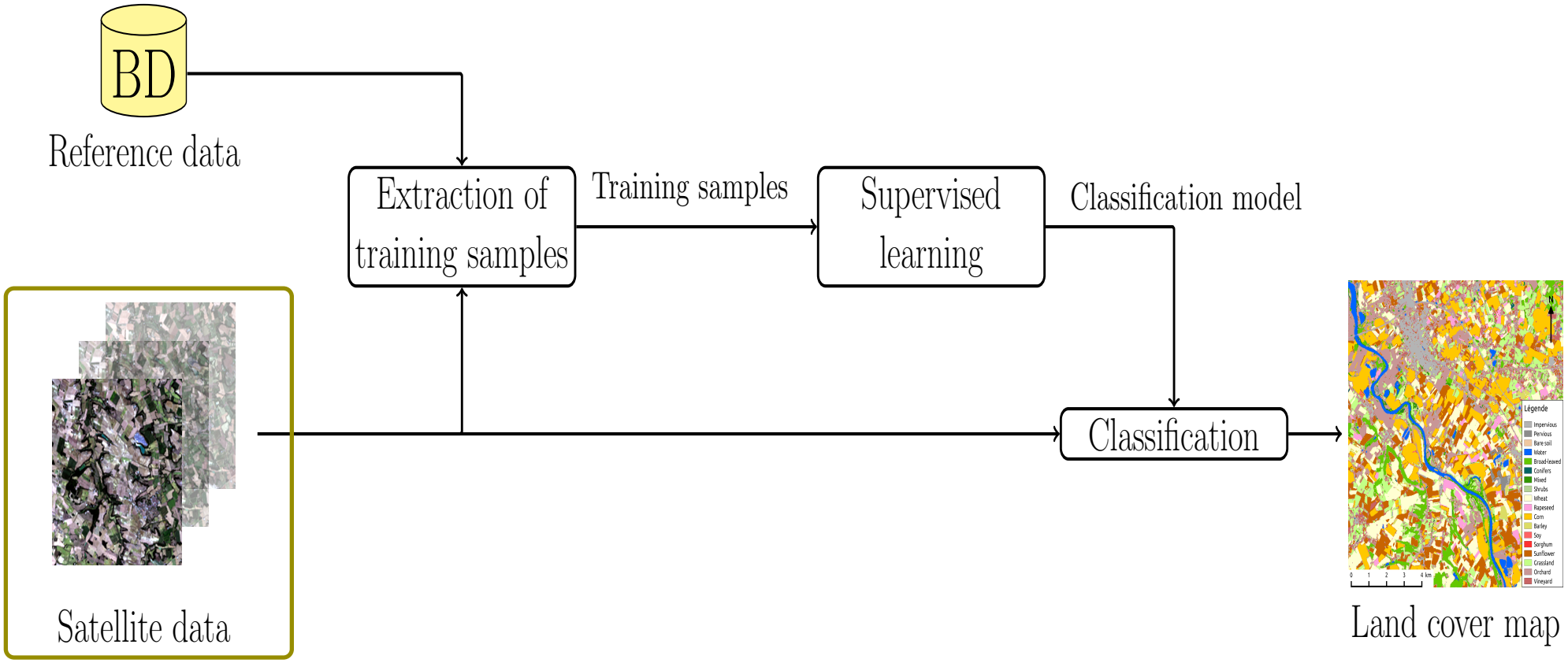
Challenges

- high volume data
- high landscape variabilities

Supervised classification framework

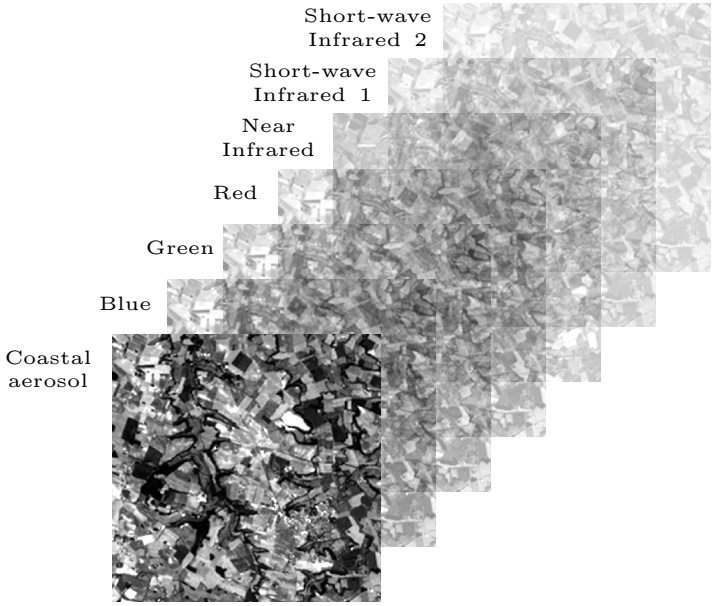


Supervised classification framework



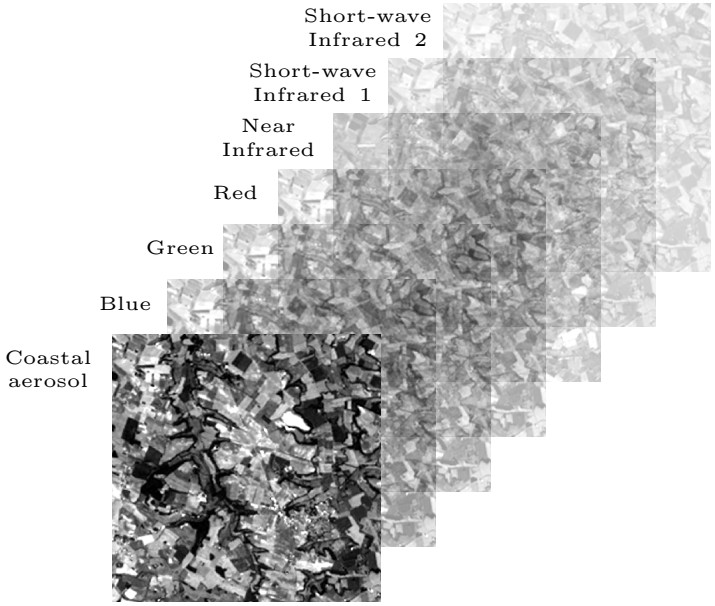
Satellite Image Time Series

Multispectral images

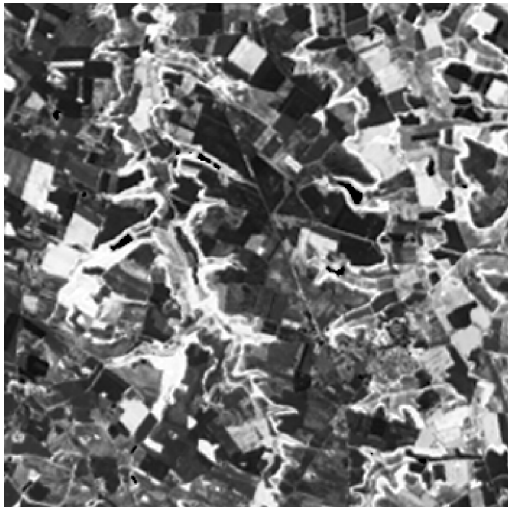


Satellite Image Time Series

Multispectral images



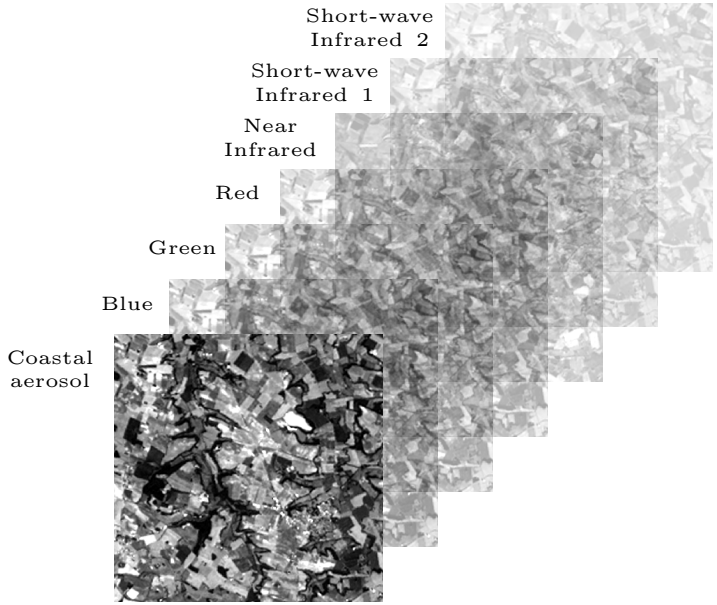
Computing spectral features



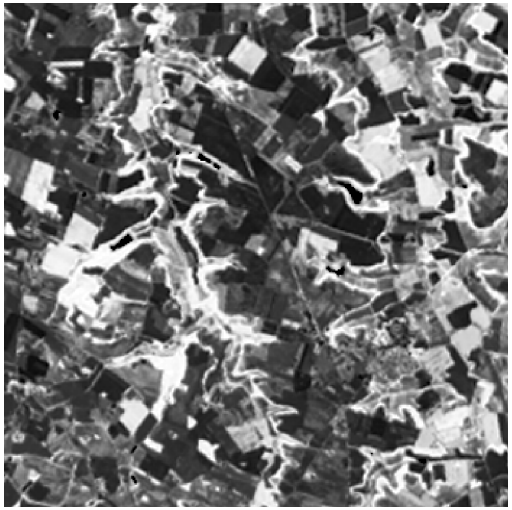
Example of NDVI (Normalized Difference Vegetation Index)

Satellite Image Time Series

Multispectral images

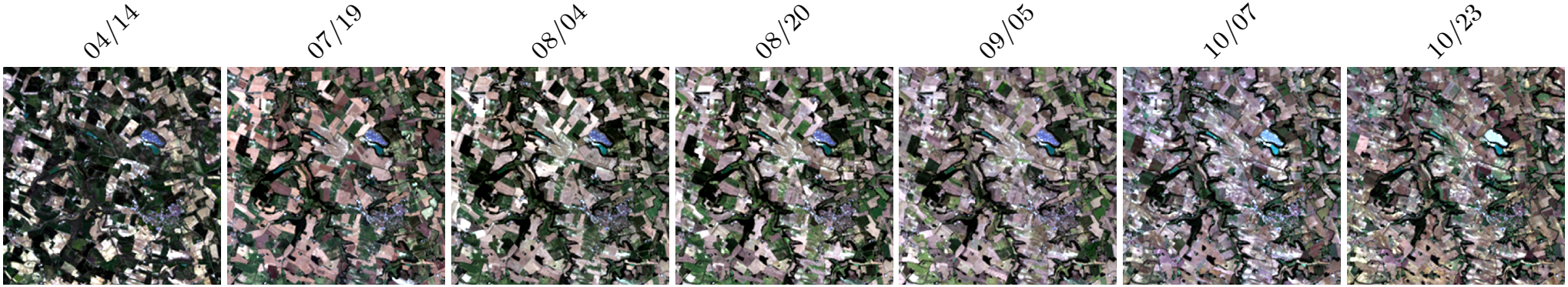


Computing spectral features

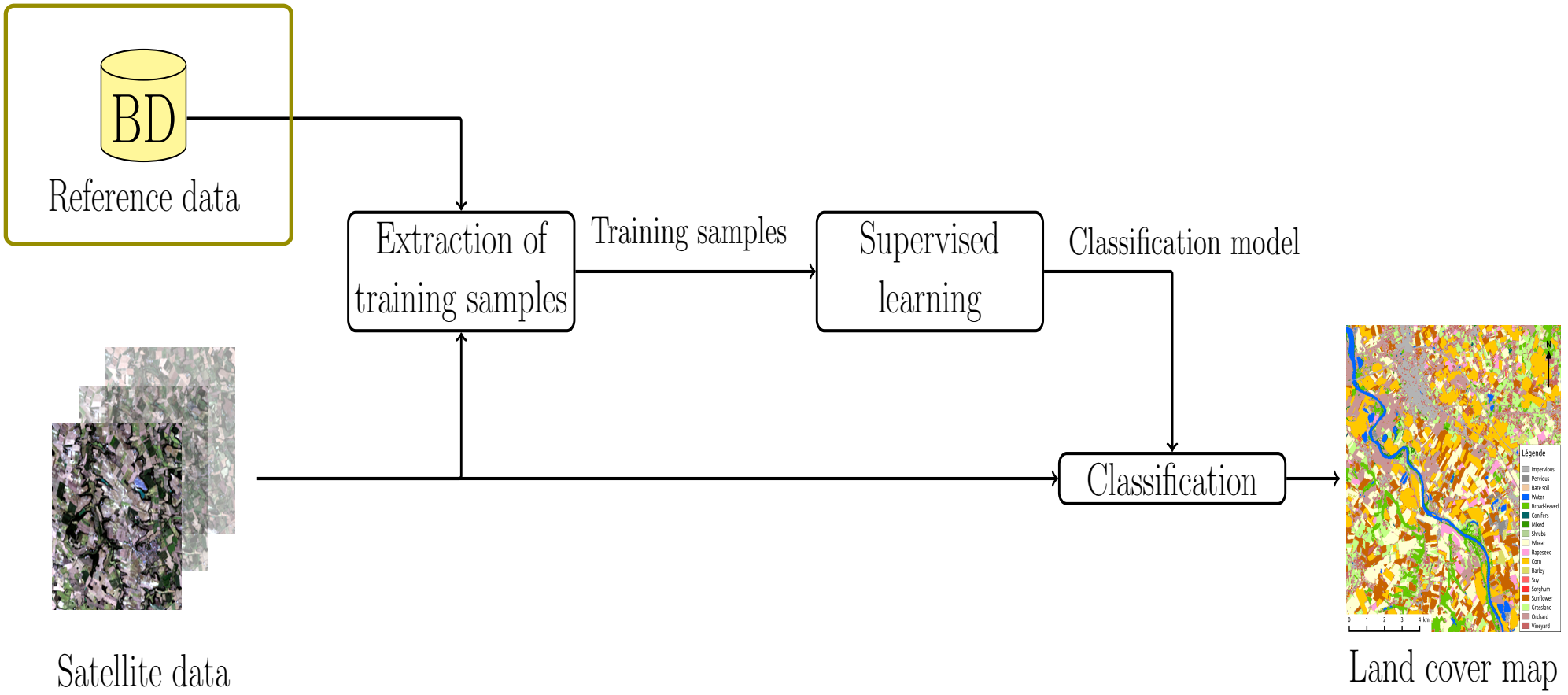


Example of NDVI (Normalized Difference Vegetation Index)

Time series



Supervised classification framework



Reference databases

Classification over large areas

- quality, accuracy
- large number of instances
- production date

Reference databases

Classification over large areas

- quality, accuracy
- large number of instances
- production date

Used reference databases

- field surveys
- photo-interpretation
- existing land cover maps
- crowd sourcing

Reference databases

Classification over large areas

- quality, accuracy
- large number of instances
- production date

Used reference databases

- field surveys
- photo-interpretation
- existing land cover maps
- crowd sourcing

Errors due to

- lack of information during field surveys
- land cover complexity
- subjectivity
- misregistration errors

Reference databases

Classification over large areas

- quality, accuracy
- large number of instances
- production date

Used reference databases

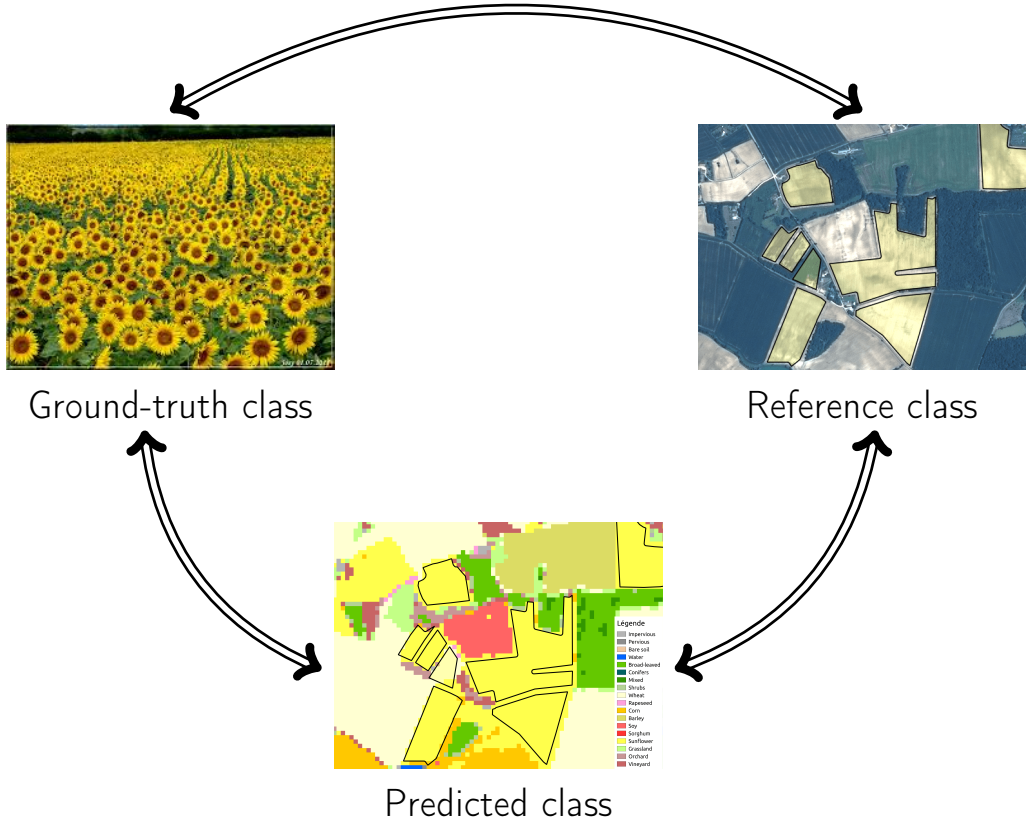
- field surveys
- photo-interpretation
- existing land cover maps
- crowd sourcing

Errors due to

- lack of information during field surveys
- land cover complexity
- subjectivity
- misregistration errors

Most adversarial errors: mislabeled data
(Zhu and Wu, 2004)

Reference databases

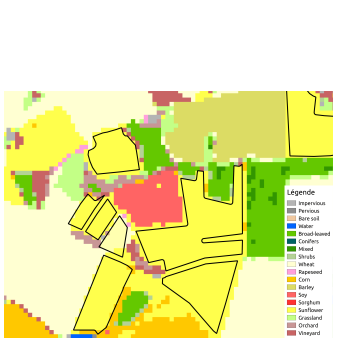


Reference databases

ground-truth $\neq$ reference

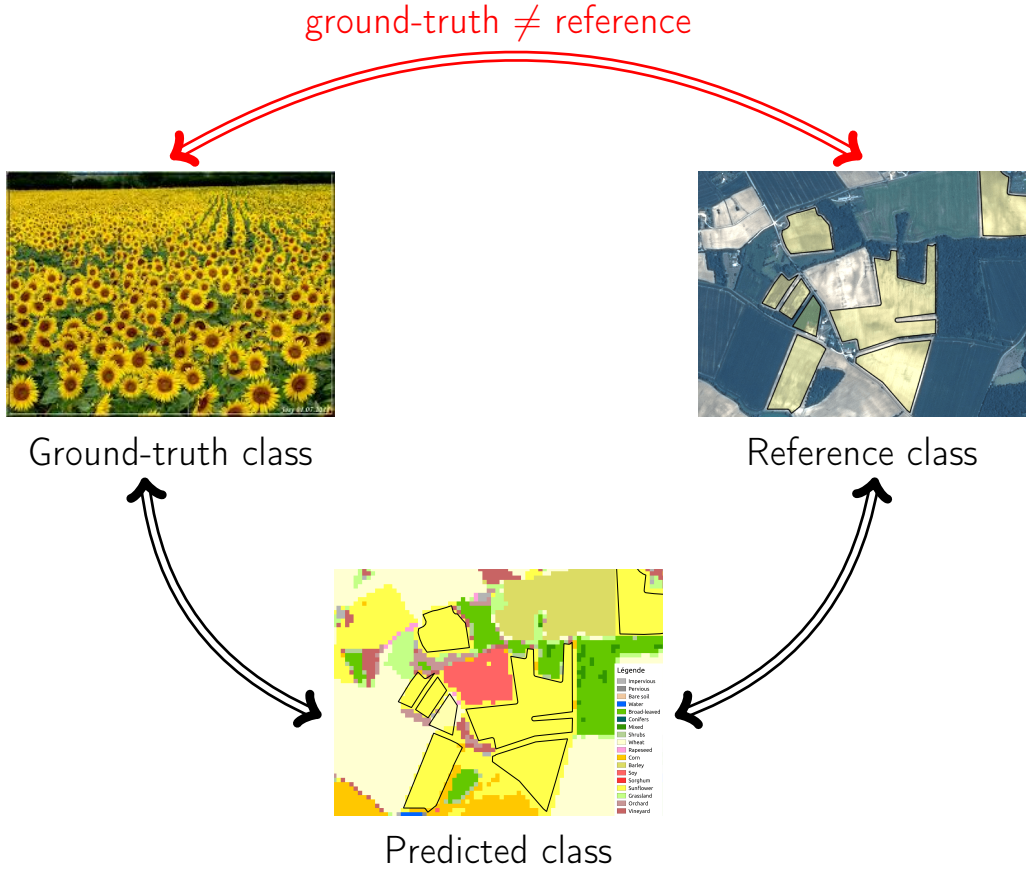
Ground-truth class

Reference class



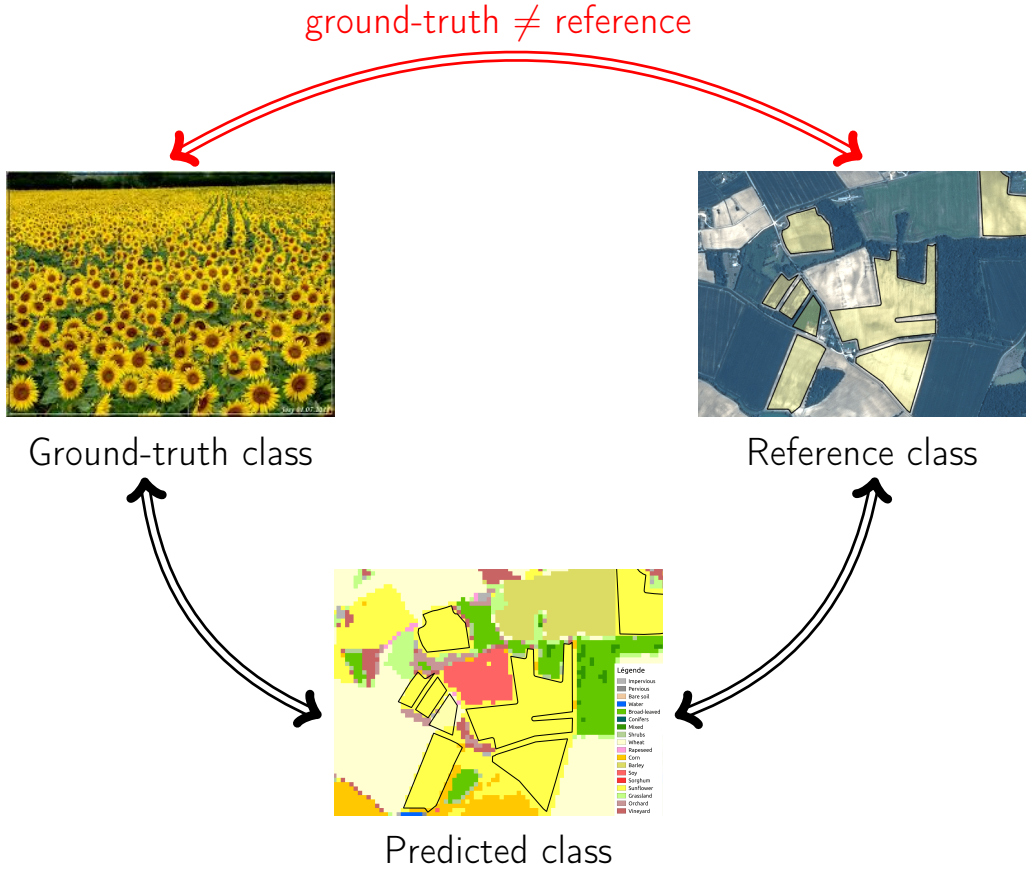
Predicted class

Reference databases



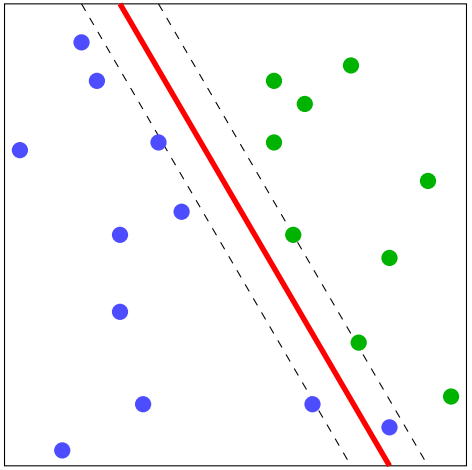
1- Consequences during validation phase
(Foody, 2002)

Reference databases



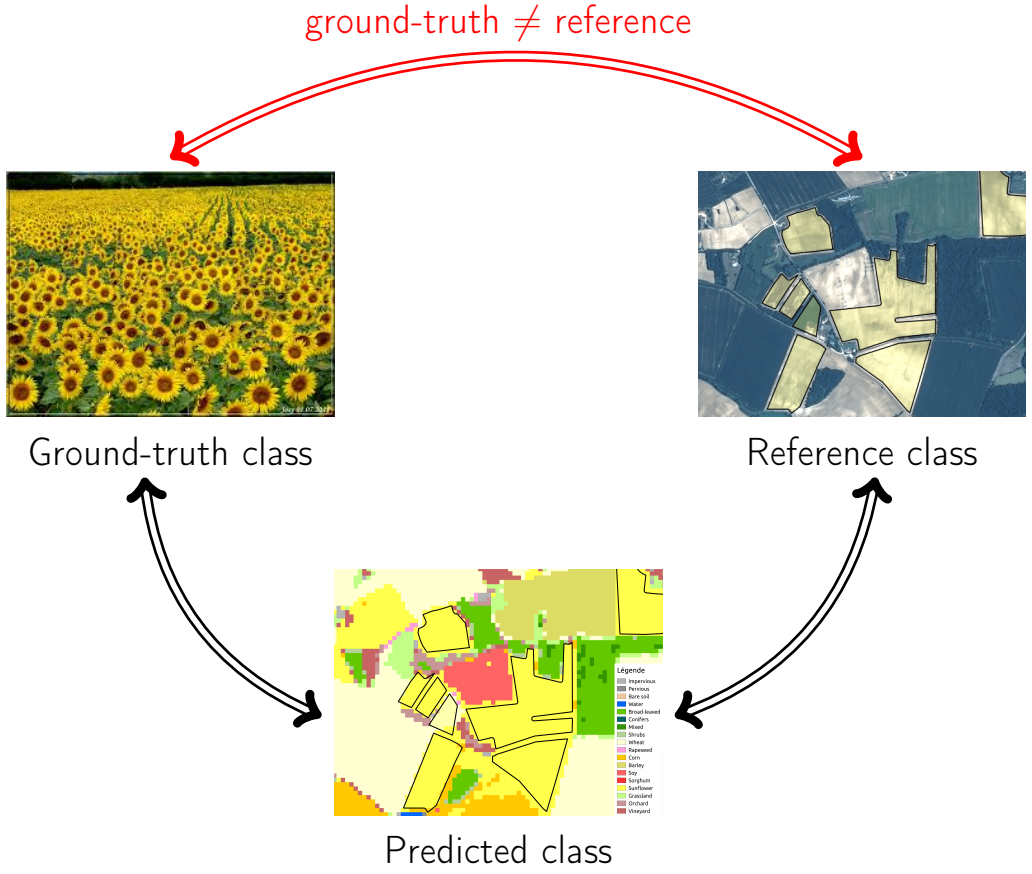
1- Consequences during validation phase (Foody, 2002)

2- **Consequences during training phase**



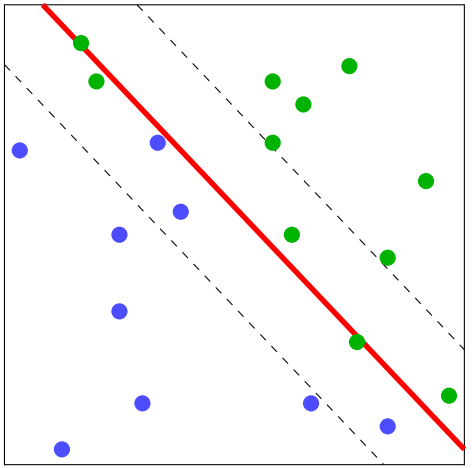
Example: SVM-Linear for 2 classes

Reference databases



1- Consequences during validation phase
(Foody, 2002)

2- **Consequences during training phase**



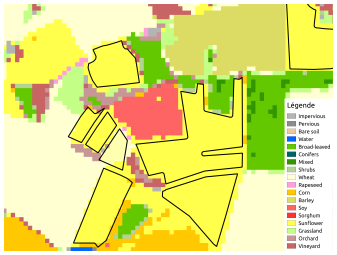
Example: SVM-Linear for 2 classes

Reference databases

ground-truth ≠ reference

Ground-truth class

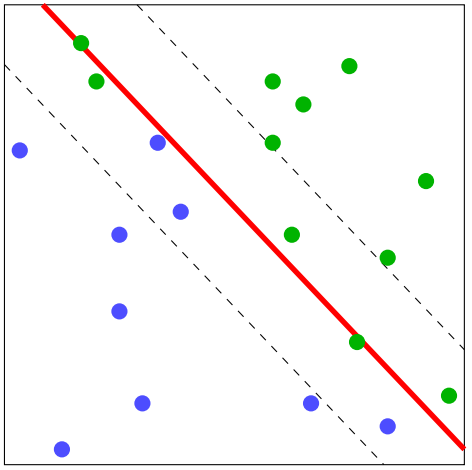
Reference class



Predicted class

1- Consequences during validation phase
(Foody, 2002)

2- **Consequences during training phase**



Example: SVM-Linear for 2 classes

Objectives

1. to study the effect of training class label noise on classification performance
2. to detect the mislabeled data
3. to propose a framework to take into account the mislabeled data

Content

Data

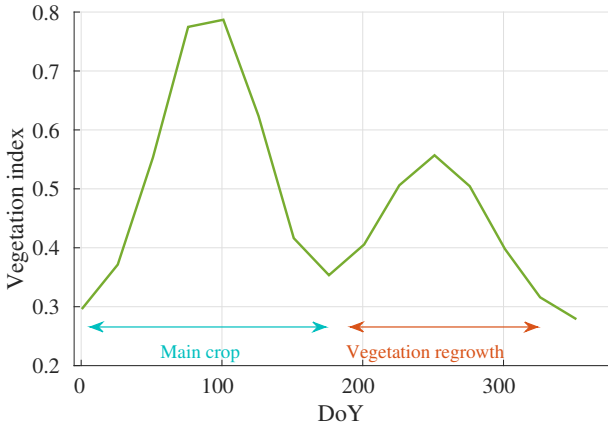
Synthetic data

- vegetation land cover classes
- synthetic NDVI profiles
- double logistic function

Data

Synthetic data

- vegetation land cover classes
- synthetic NDVI profiles
- double logistic function

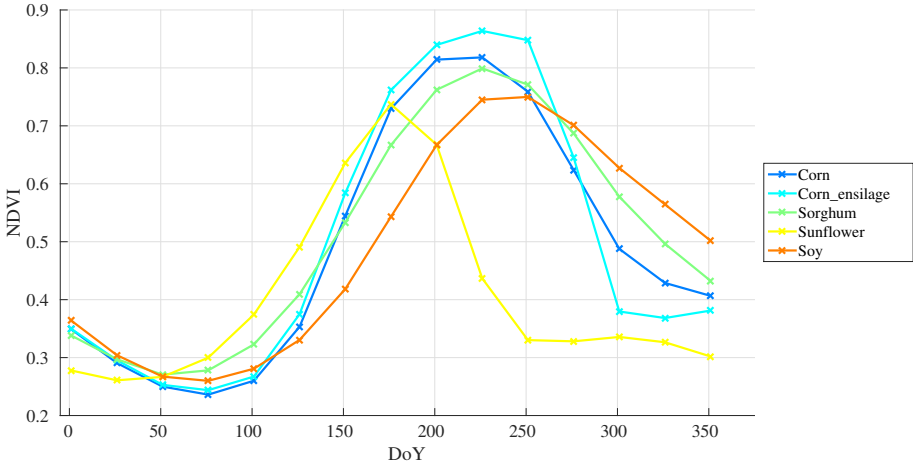


Data

Synthetic data

- vegetation land cover classes
- synthetic NDVI profiles
- double logistic function

- the level of noise is known



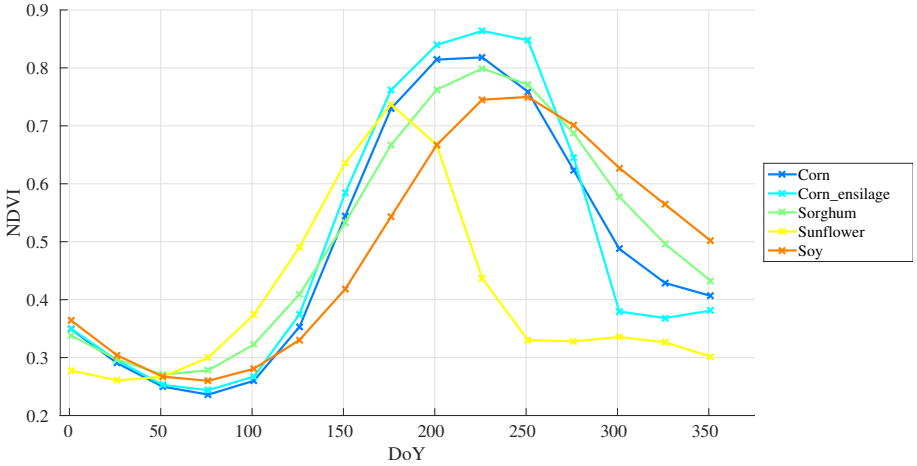
15 dates

Data

Synthetic data

- vegetation land cover classes
- synthetic NDVI profiles
- double logistic function

- the level of noise is known



15 dates

Real data

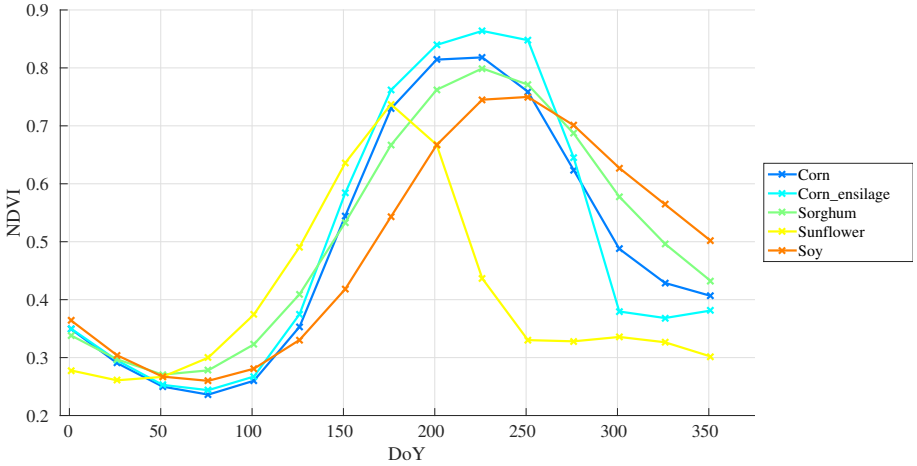
- vegetation land cover classes
- real NDVI profiles
- satellite image time series

Data

Synthetic data

- vegetation land cover classes
- synthetic NDVI profiles
- double logistic function

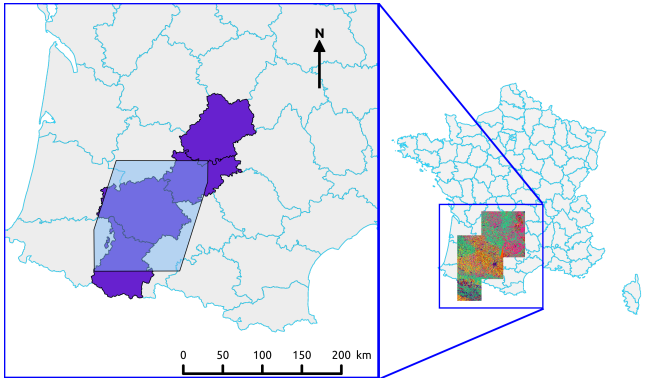
- the level of noise is known



15 dates

Real data

- vegetation land cover classes
- real NDVI profiles
- satellite image time series



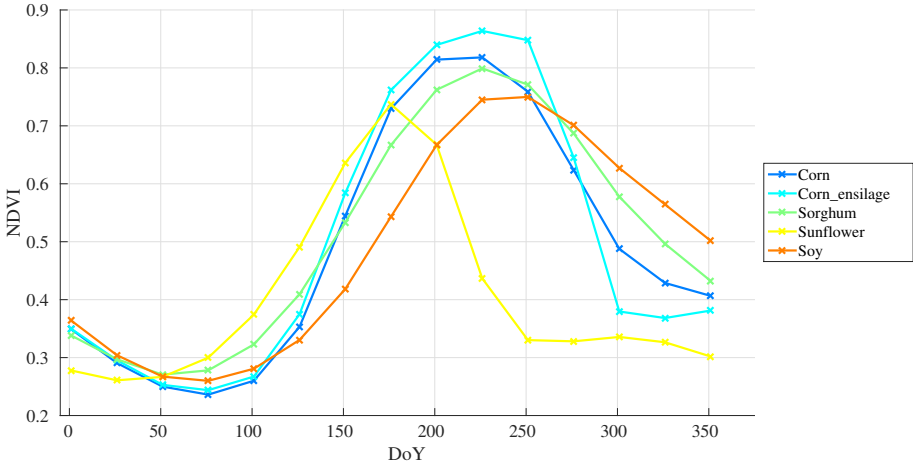
Landsat-8 and SPOT 4 images – farmer's declaration 2013

Data

Synthetic data

- vegetation land cover classes
- synthetic NDVI profiles
- double logistic function

- the level of noise is known

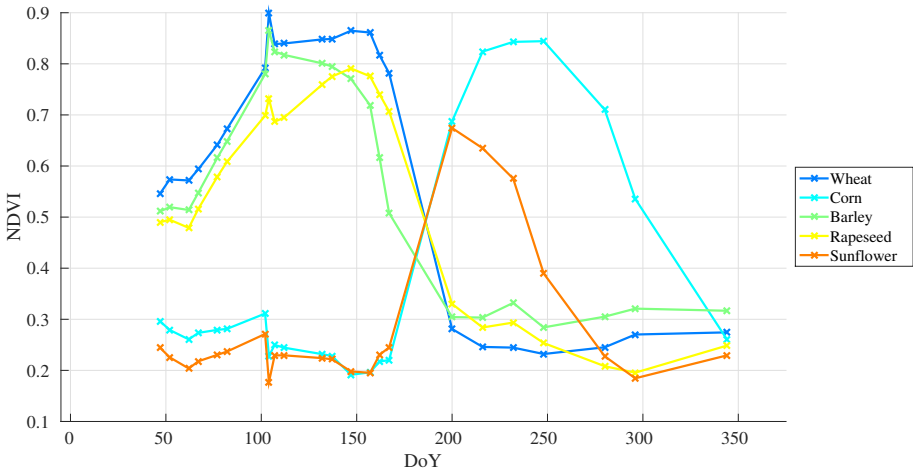


15 dates

Real data

- vegetation land cover classes
- real NDVI profiles
- satellite image time series

- the landscape complexity is taken into account



23 dates

Data

Datasets	2-class dataset	5-class dataset	10-class dataset
Synthetic	Corn Corn silage	Corn Corn silage Sorghum Sunflower Soy	Corn Corn silage Sorghum Sunflower Soy Wheat Rapeseed Barley Evergreen Deciduous
Real	Corn Sunflower	Corn Sunflower Wheat Barley Rapeseed	-

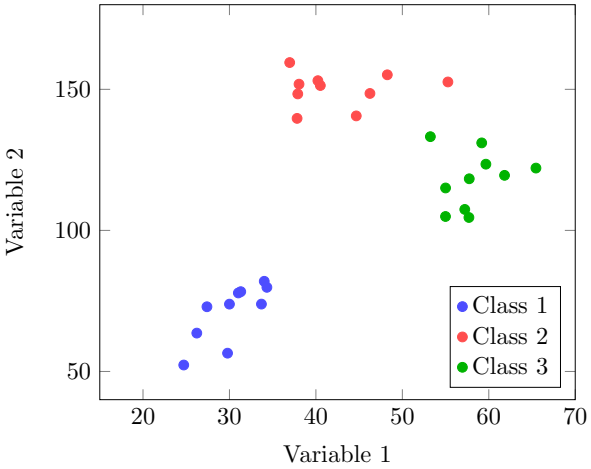
- Number of instances per class
  - 500 for training (50 polygons of 10 instances)
  - 500 for testing (50 polygons of 10 instances)
- 10 random trials

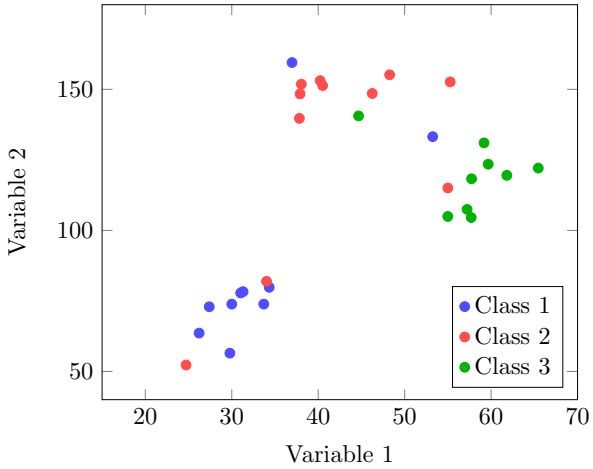
Noise generation procedure

State-of-the-art

(Zhu and Wu, 2003)

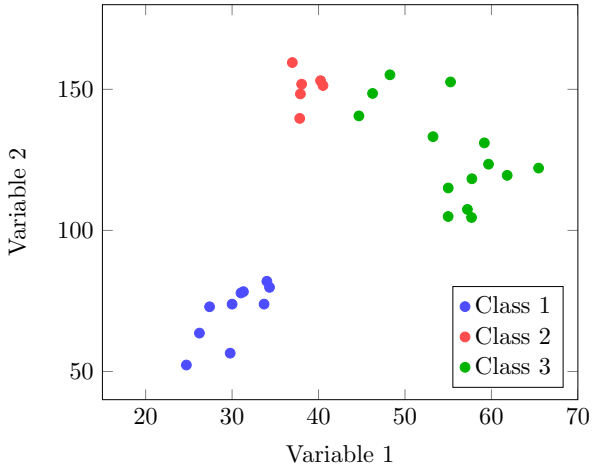
- Random noise
- Rule-based noise
  - expert's knowledge
  - common confusions





⊕ easy to generate

⊖ unrealistic



⊕ realistic

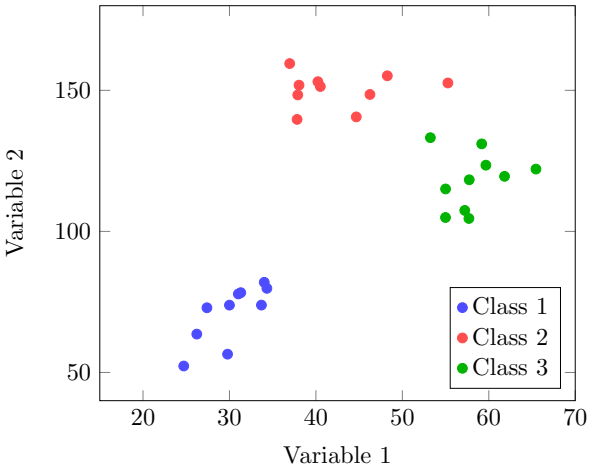
⊖ subjectivity

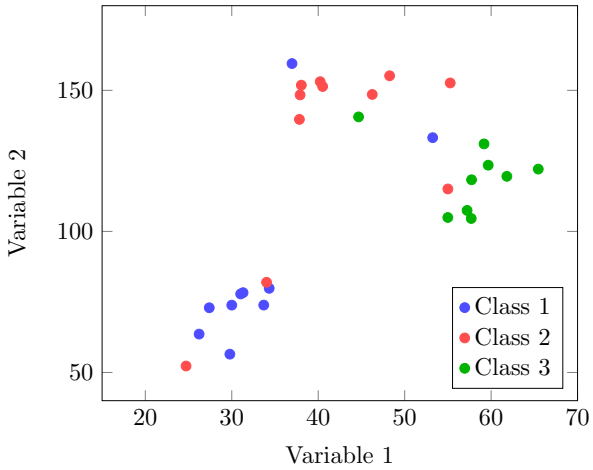
Noise generation procedure

State-of-the-art

(Zhu and Wu, 2003)

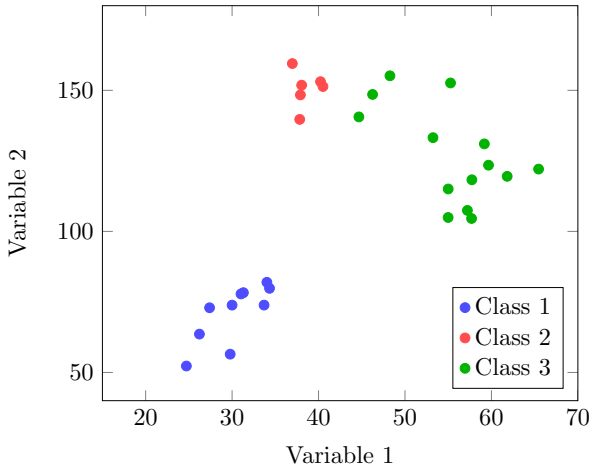
- Random noise
- Rule-based noise
  - expert's knowledge
  - common confusions





⊕ easy to generate

⊖ unrealistic



⊕ realistic

⊖ subjectivity

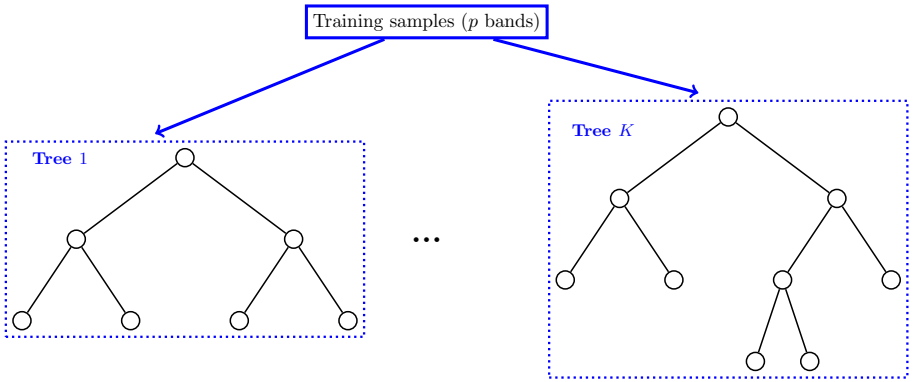
Protocol

- 20 levels of noise from 5 to 100 % by 5 % step
- by polygon
- the class label noise affects only training instances

Studied classifiers

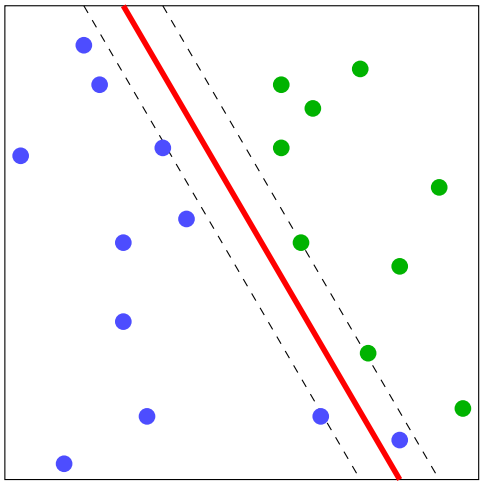
Classical classifiers for remote sensing data (Khatami *et al.*, 2016)

Random Forests (RF)
(Breiman, 2001)



- Parameters
  - number of trees = 100
  - number of features randomly selected at each node = $\sqrt{\text{number of features}}$

Support Vector Machines (SVM)
(Vapnik, 1995)

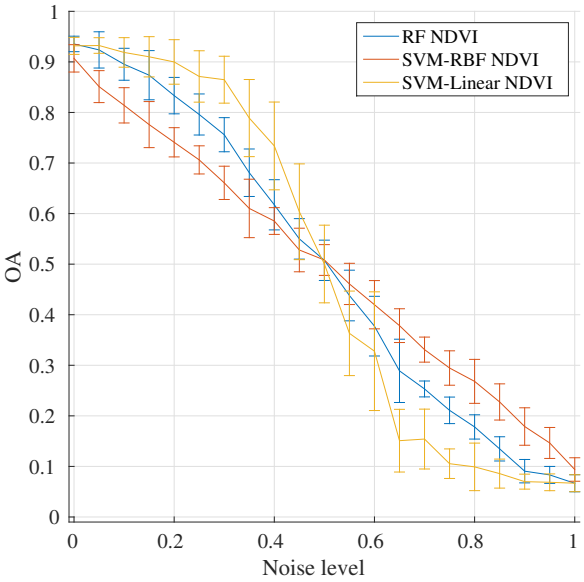


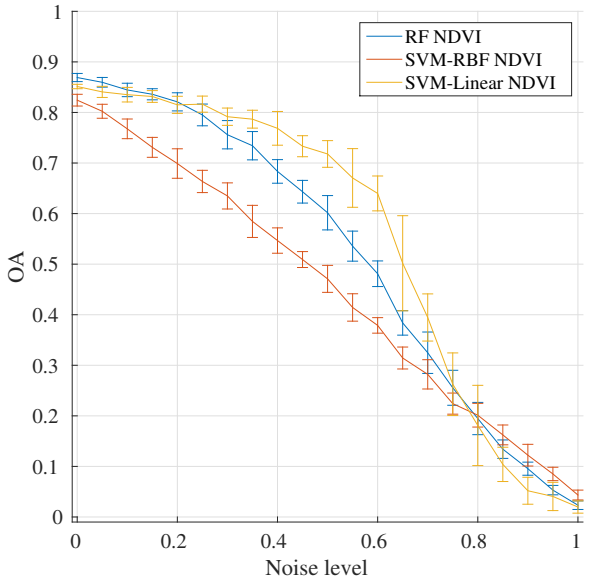
- Kernel parameters
  - linear C
  - gaussian C and γ
- Optimization by 5-fold cross validation

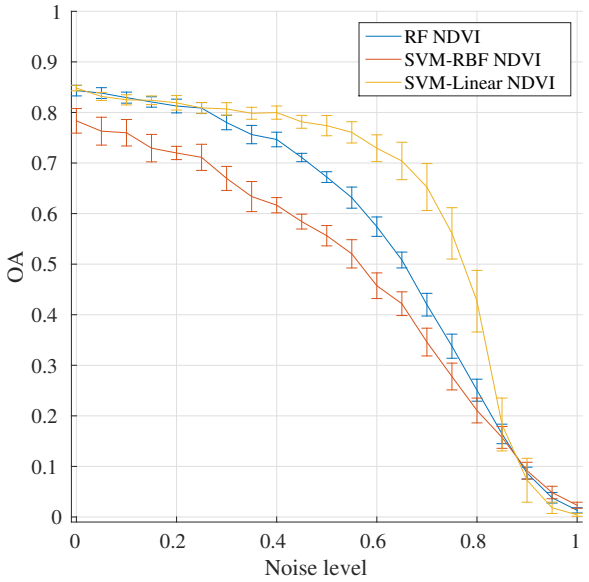
1st study: Influence of the number of classes

2 classes | 5 classes | 10 classes

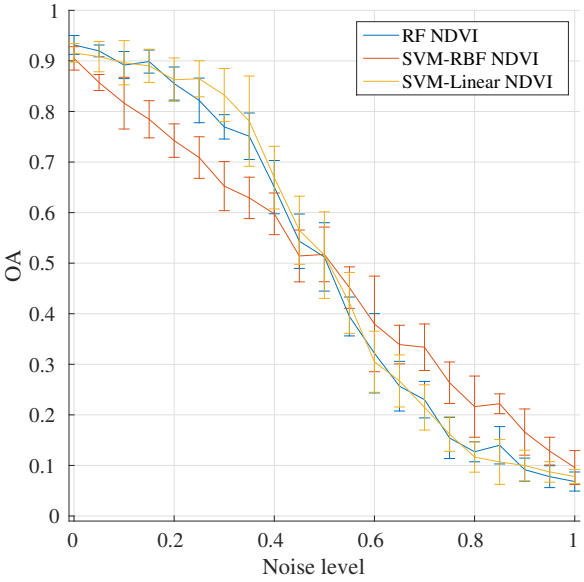
Synthetic

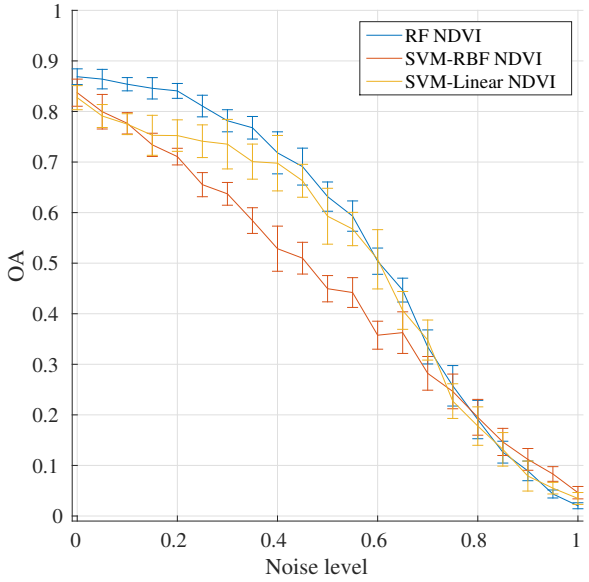






Real

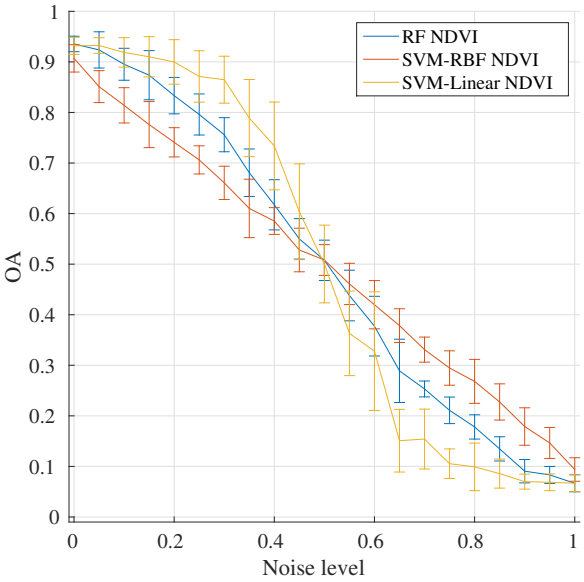


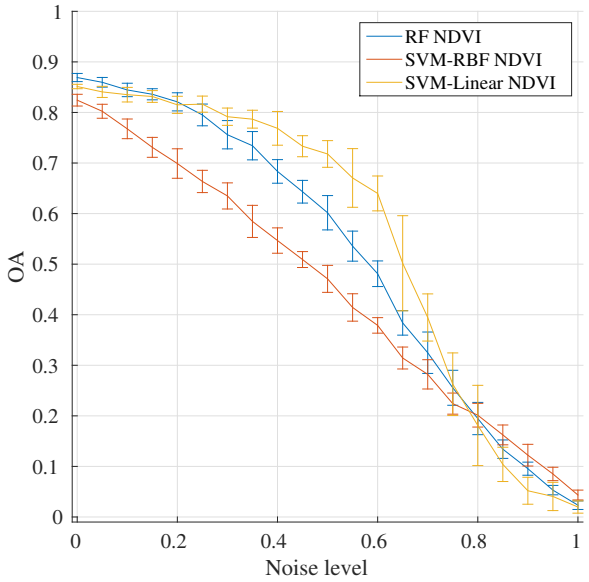


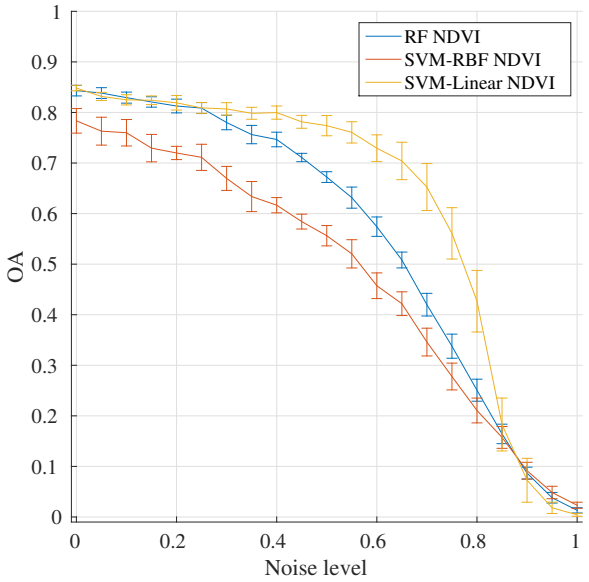
1st study: Influence of the number of classes

2 classes | 5 classes | 10 classes

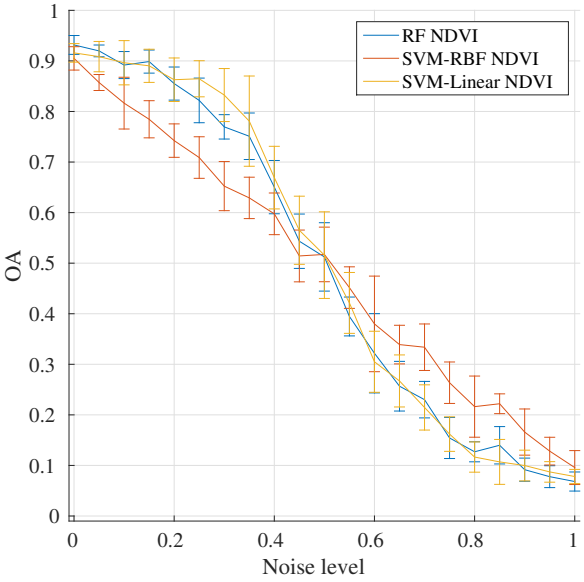
Synthetic

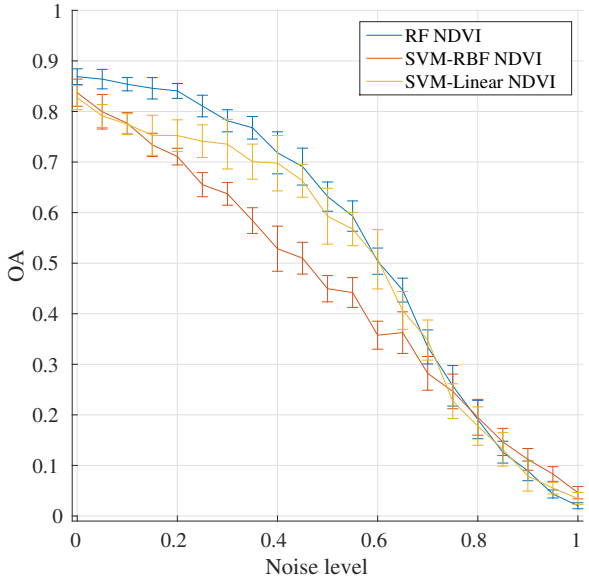






Real

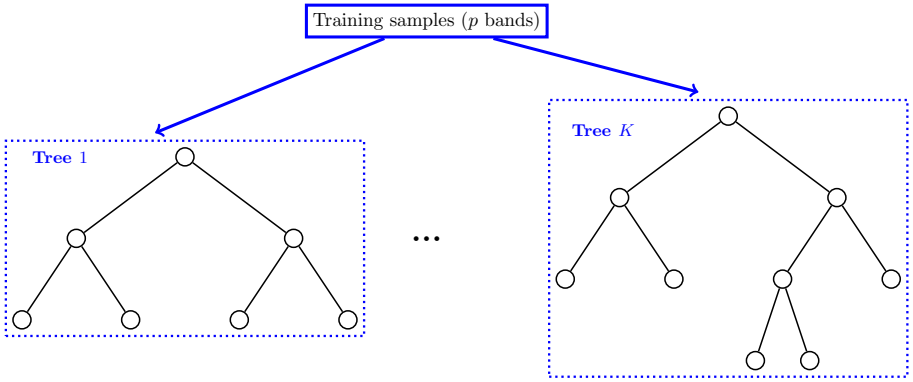




- SVM-Linear > RF > SVM-RBF
- nb. classes ↗ ⇒ OA ↘
- nb. classes ↗ ⇒ flat zone length ↗
- exception for 5-class real dataset

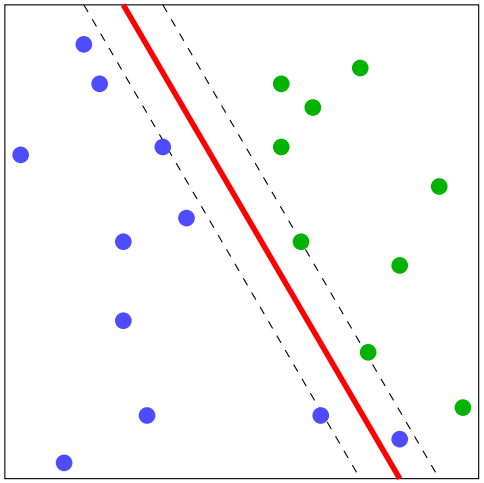
2nd study: Algorithm complexity

Random Forests



- Average path length of each training samples

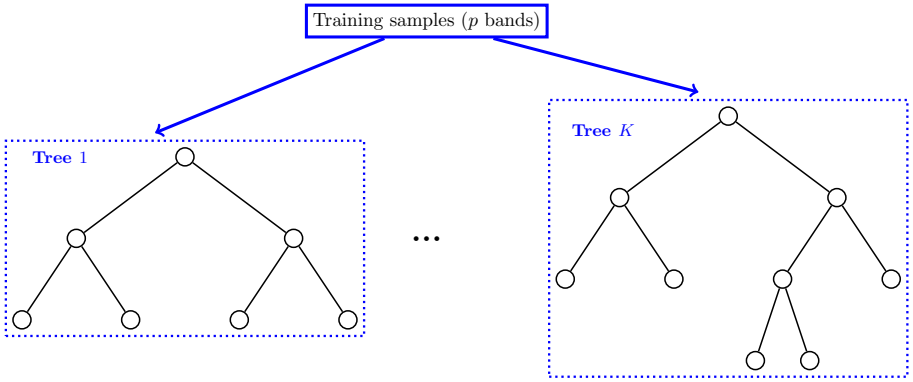
Support Vector Machines



- Number of support vectors

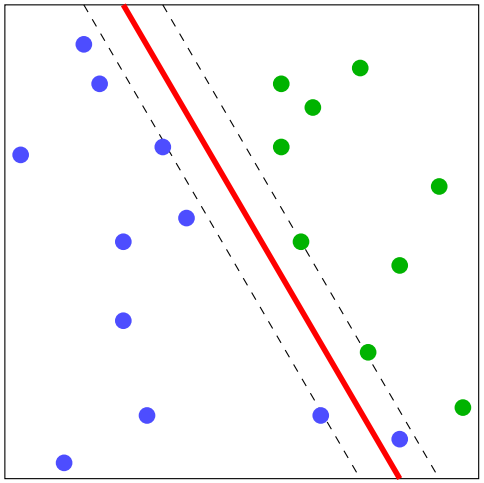
2nd study: Algorithm complexity

Random Forests



- Average path length of each training samples

Support Vector Machines



- Number of support vectors

Protocol

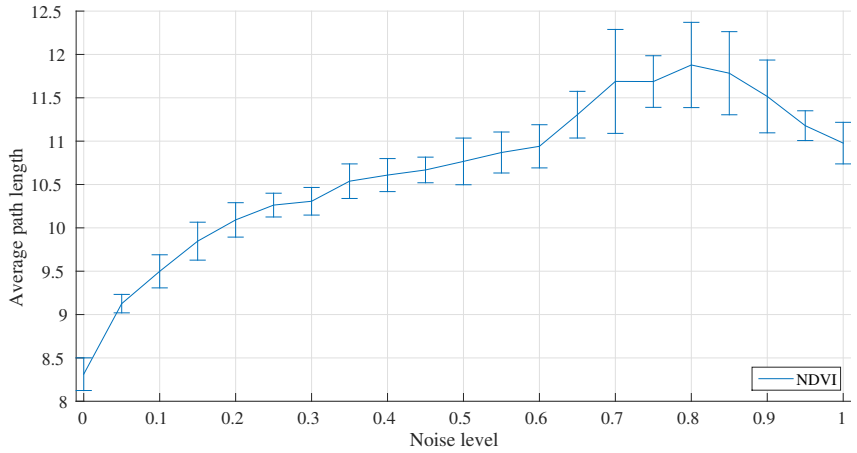
- 5 classes for real and synthetic datasets

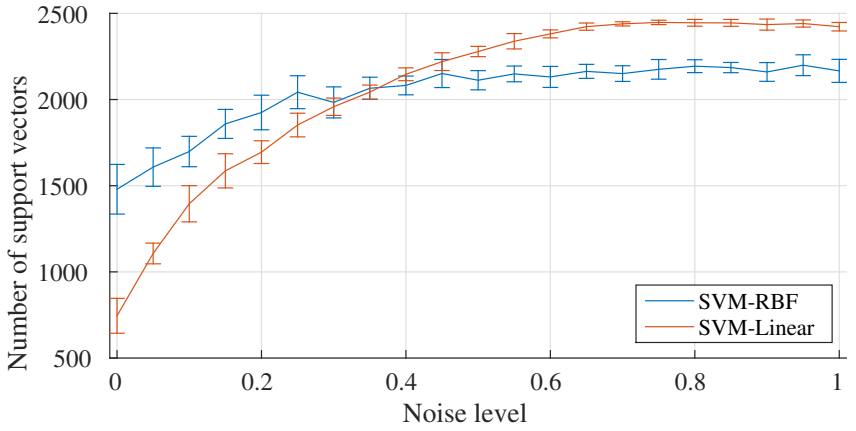
2nd study: Algorithm complexity

RF

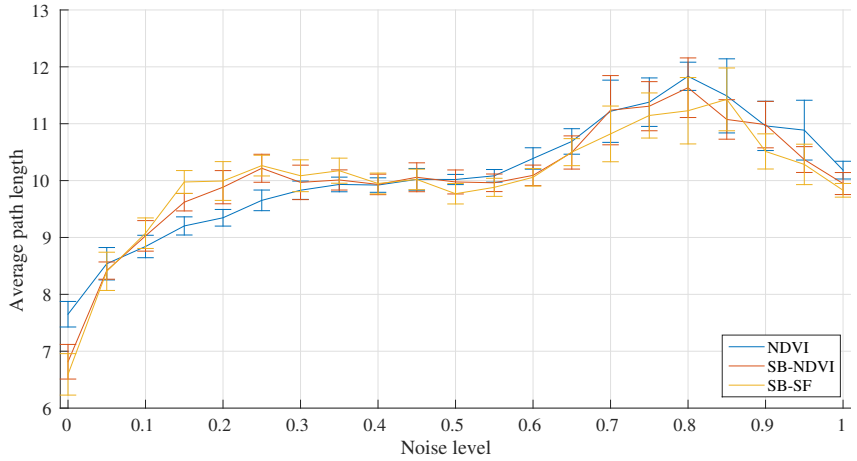
SVM

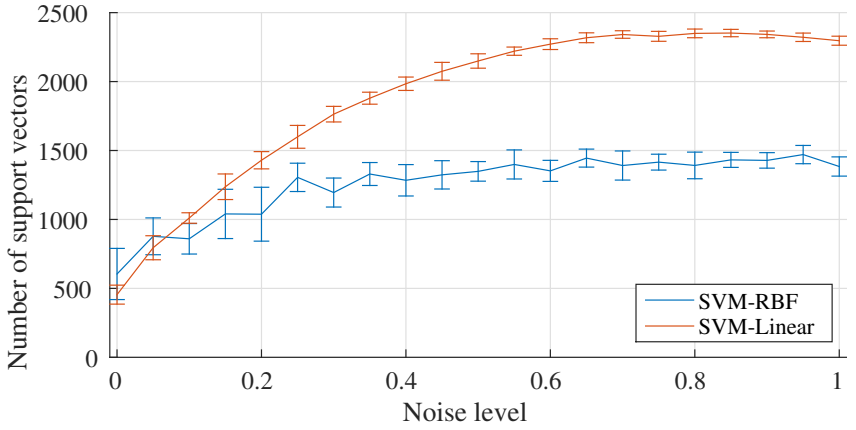
Synthetic
5 classes





Real
5 classes





Conclusions

- mislabeled training data influence the classification performance
- RF more robust than SVM classifiers (more studies with different configurations)
- OA values are stable for low noise levels and OA drops for higher noise levels
- sensitivity of SVM classifier to its parameter setting
- algorithm complexity $\nearrow$ with noise level

Conclusions

- mislabeled training data influence the classification performance
- RF more robust than SVM classifiers (more studies with different configurations)
- OA values are stable for low noise levels and OA drops for higher noise levels
- sensitivity of SVM classifier to its parameter setting
- algorithm complexity $\nearrow$ with noise level

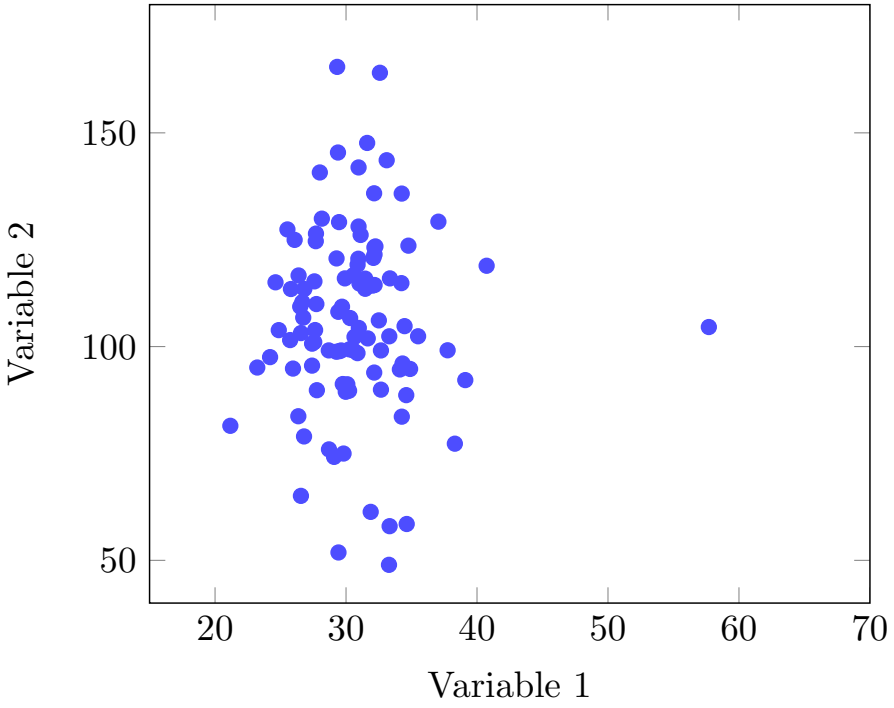
$\Rightarrow$ **Detection of mislabeled training data**

Content

Outliers

Définition :

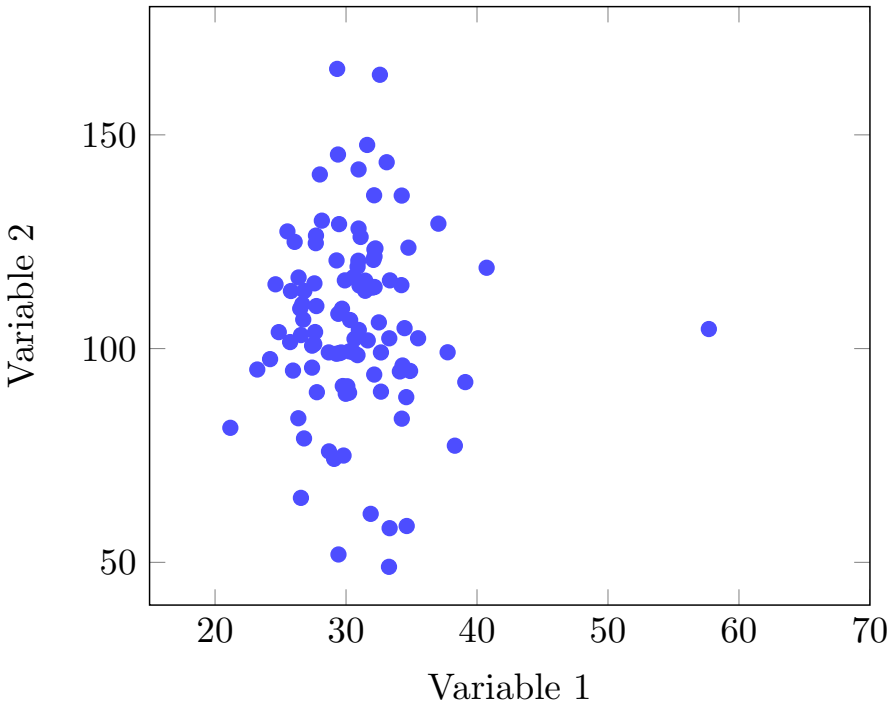
> *An outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism*
>
> (Hawkins, 1980)



Outliers

Définition :

> *An outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism*
>
> (Hawkins, 1980)

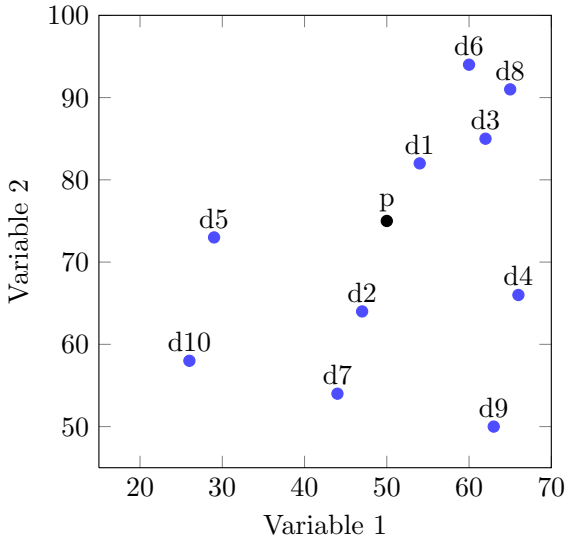


$\Rightarrow$ closely related to the notion of the mislabeled

Detecting mislabeled data

Literature

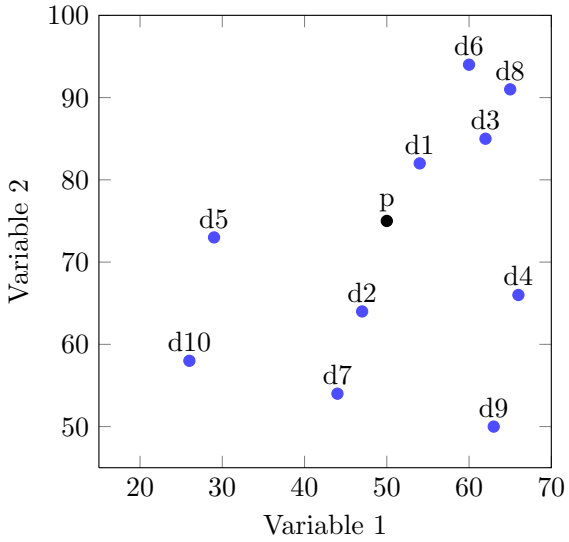
1. computing an outlier score
   for the instance p



Detecting mislabeled data

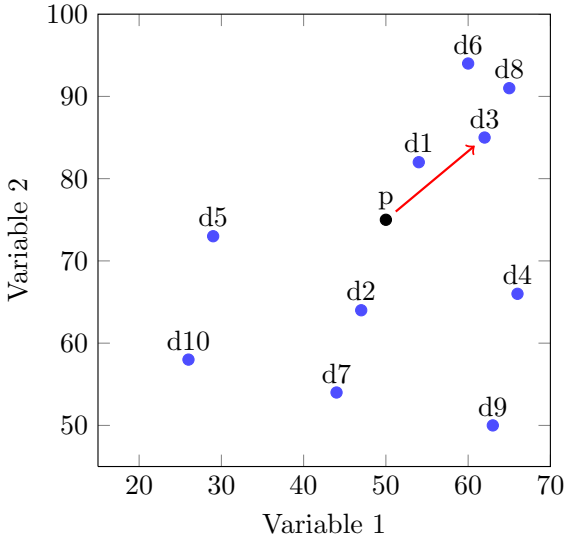
Literature

1. computing an outlier score for the instance p



k-Nearest Neighbor (kNN)

(Ramaswamy *et al.*, 2000)

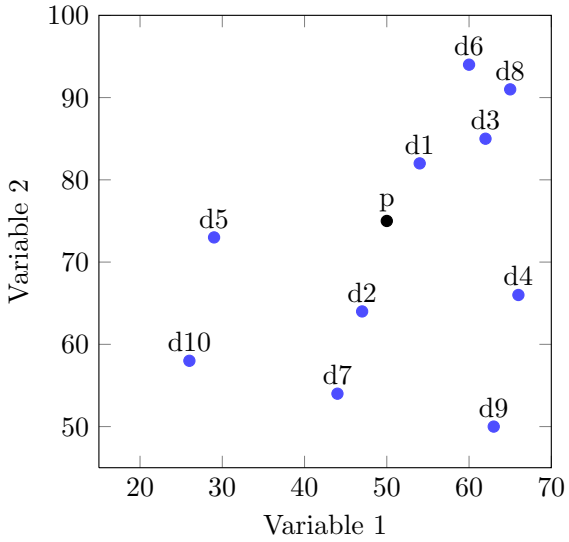


outlier score is the distance to the kth nearest neighbor ($k = 3$)

Detecting mislabeled data

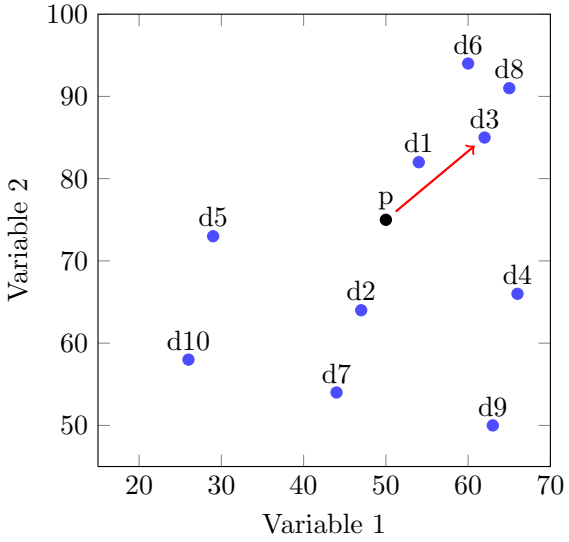
Literature

1. computing an outlier score for the instance p



k-Nearest Neighbor (kNN)

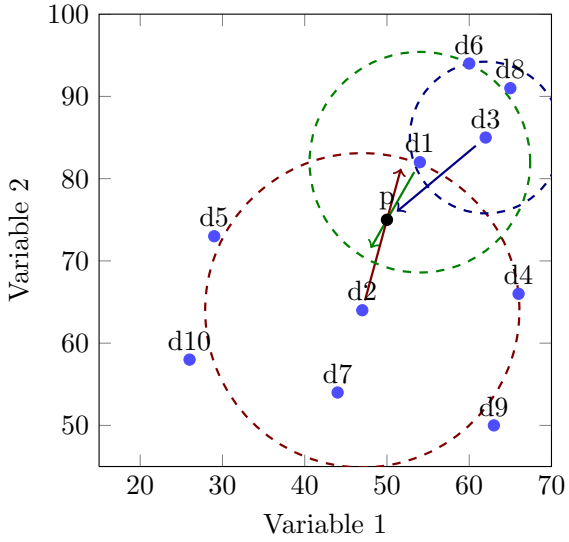
(Ramaswamy *et al.*, 2000)



outlier score is the distance to the kth nearest neighbor ($k = 3$)

Local Outlier Factor (LOF)

(Breuning *et al.* 2000)

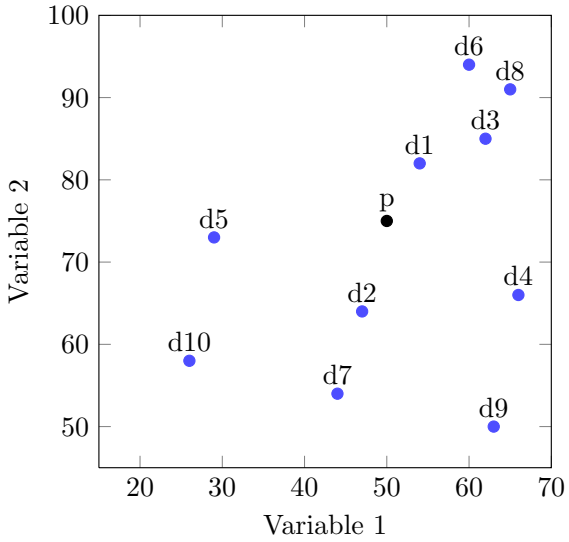


similar to kNN, but LOF computes local densities

Detecting mislabeled data

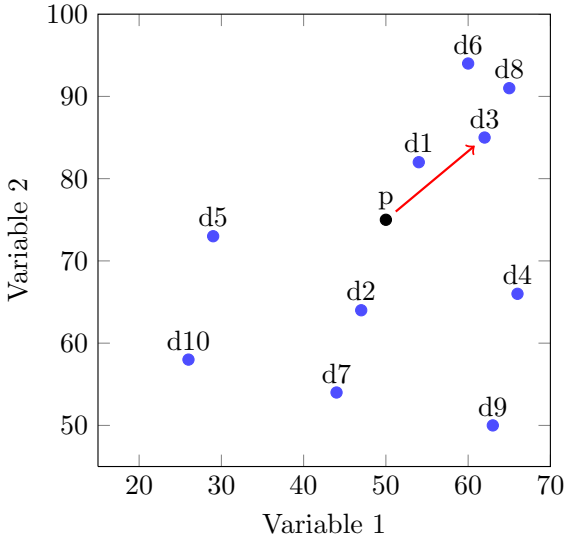
Literature

1. computing an outlier score for the instance p



k-Nearest Neighbor (kNN)

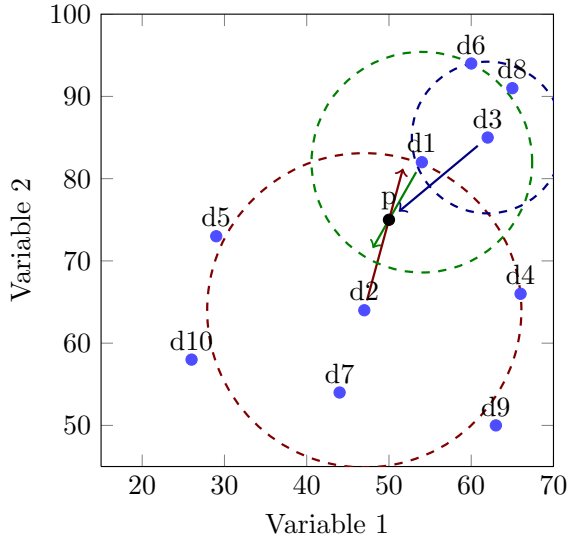
(Ramaswamy *et al.*, 2000)



outlier score is the distance to the kth nearest neighbor ($k = 3$)

Local Outlier Factor (LOF)

(Breuning *et al.* 2000)



similar to kNN, but LOF computes local densities

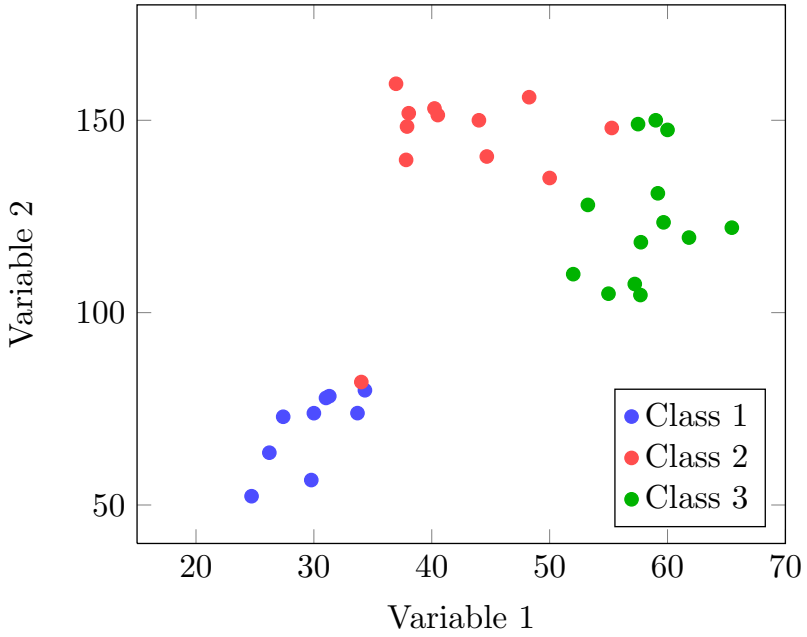
Drawbacks of these approaches

- k setting
- Euclidean distance

Detecting mislabeled data

Literature

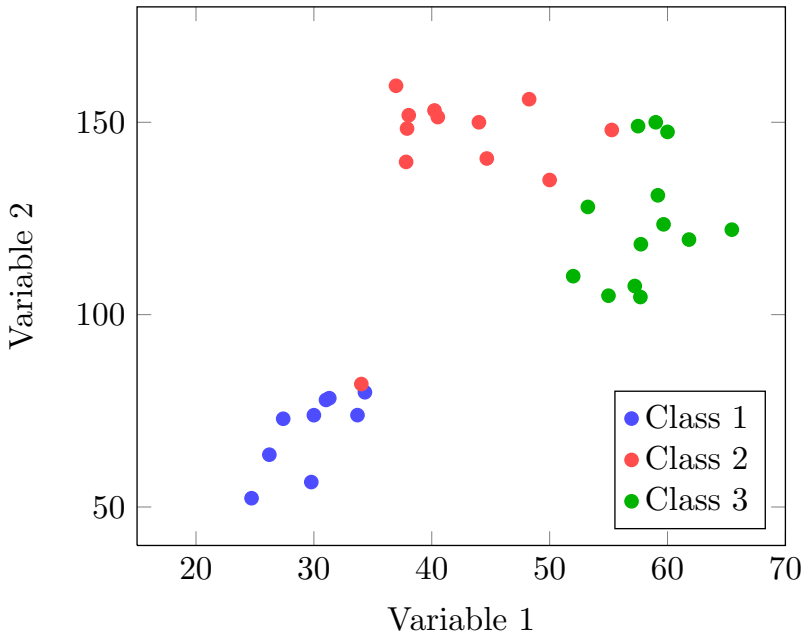
2. using editing method



Detecting mislabeled data

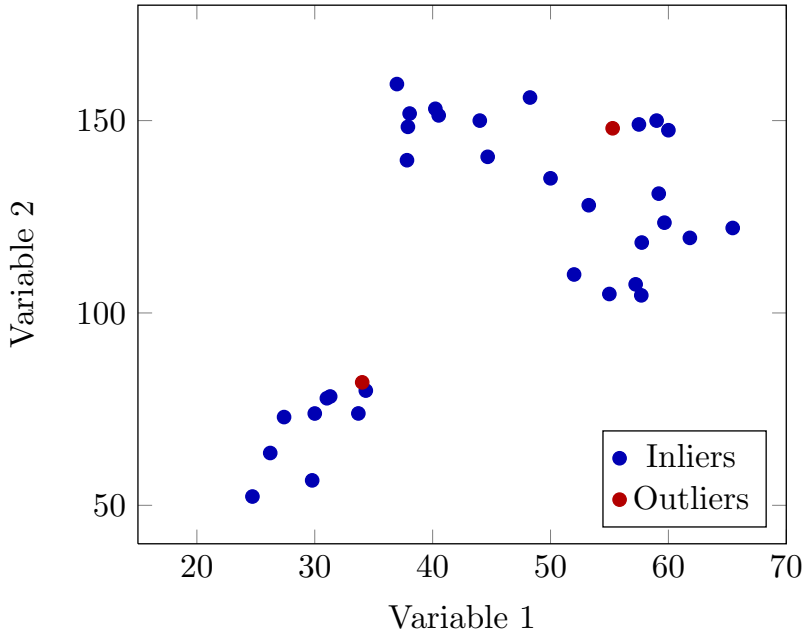
Literature

2. using editing method



Edited Nearest Neighbor (ENN)

(Wilson, 1972)

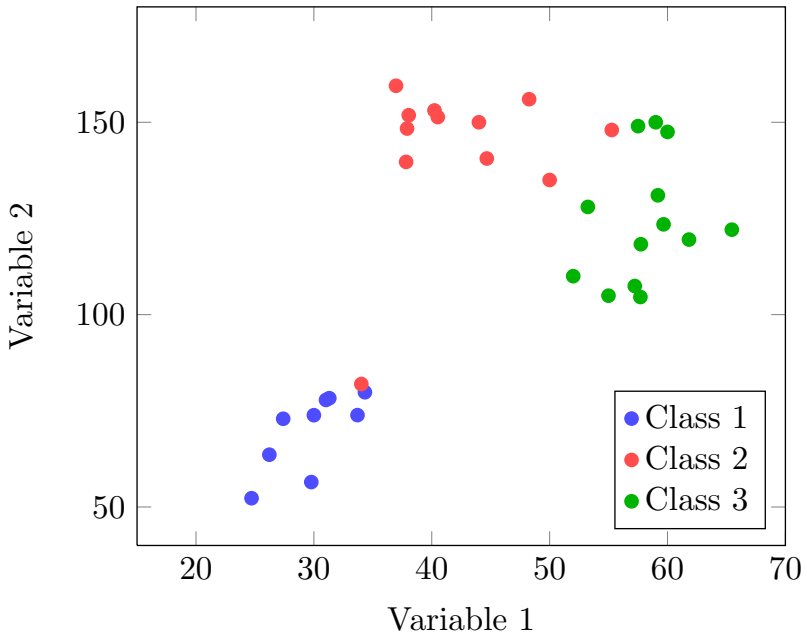


an instance is an outlier if its k nearest neighbor have a different label ($k = 3$)

Detecting mislabeled data

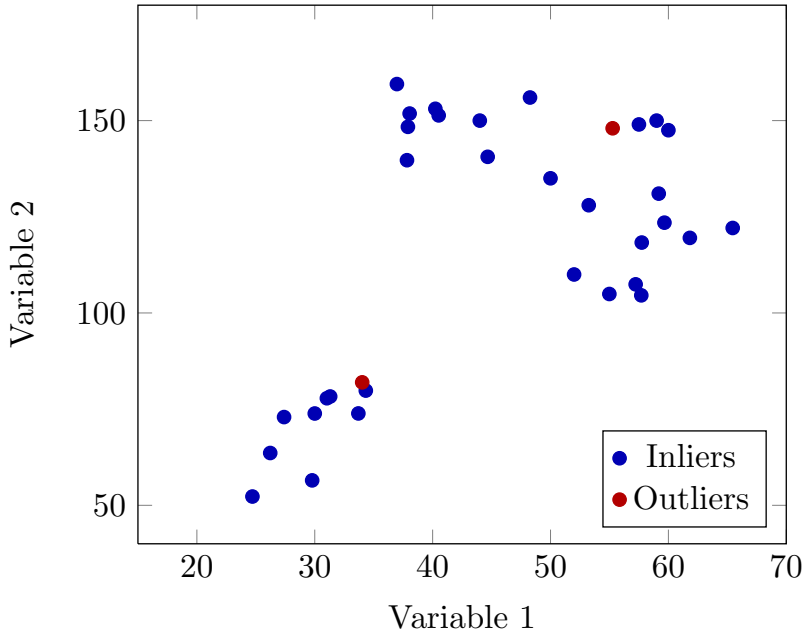
Literature

2. using editing method



Edited Nearest Neighbor (ENN)

(Wilson, 1972)



an instance is an outlier if its k nearest neighbor have a different label ($k = 3$)

Drawbacks of this approach

- k setting
- Euclidean distance
- difficult to apply for imbalanced classification problemn

Detecting mislabeled data

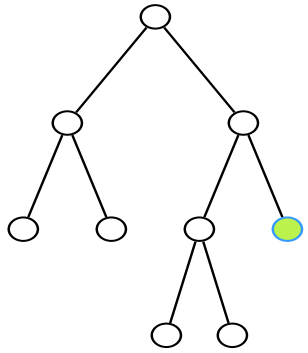
3. computing Random Forest outlier score
   - Two steps are required
     a. computing proximities
     b. converting proximities into outlier scores

Detecting mislabeled data

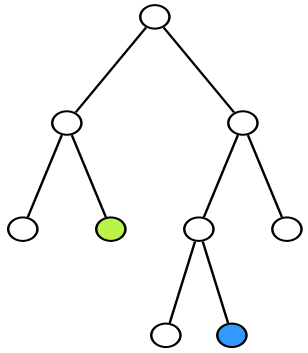
3. computing Random Forest outlier score
   - Two steps are required
     a. computing proximities
     b. converting proximities into outlier scores

a. Computing proximities between each pair of instances

- Using RF tree structure
- For each tree
  - +1 if the instances fall into the same terminal node
  - +0 otherwise
- Divided by the number of trees
  $\Rightarrow$ proximity between i and j: $0 \leq P(i,j) \leq 1$



+1



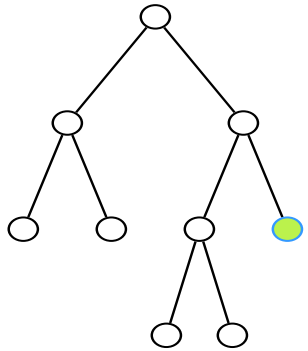
+0

Detecting mislabeled data

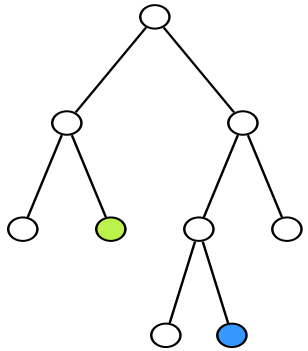
3. computing Random Forest outlier score
   - Two steps are required
     a. computing proximities
     b. converting proximities into outlier scores

a. Computing proximities between each pair of instances

- Using RF tree structure
- For each tree
  - +1 if the instances fall into the same terminal node
  - +0 otherwise
- Divided by the number of trees
  $\Rightarrow$ proximity between i and j: $0 \leq P(i,j) \leq 1$



+1

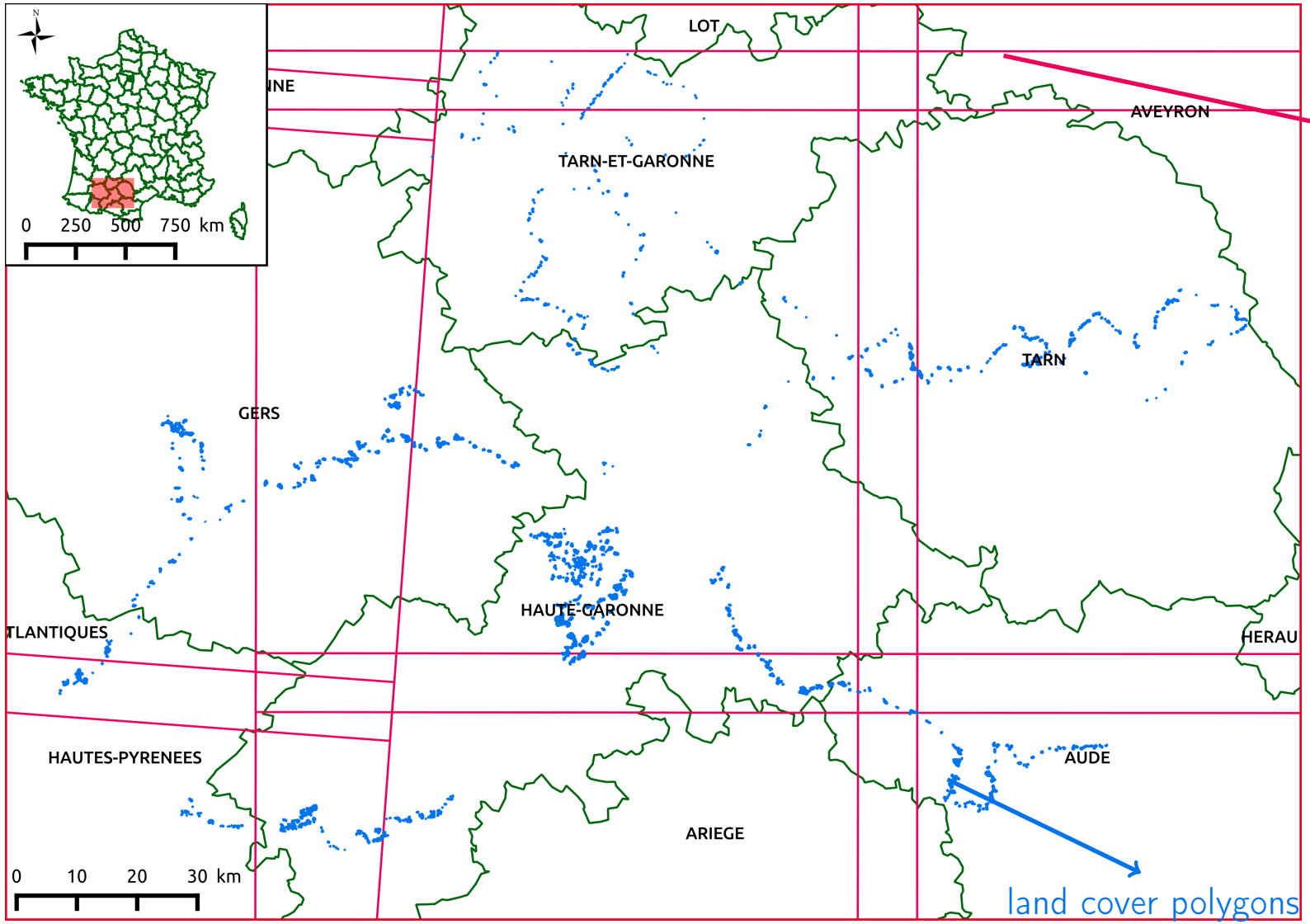


+0

b. Converting proximities into outlier scores

- outliers are the instances having small proximities to all other instances
- outlier score: $OS(i) \sim \frac{1}{\sum_j P(i,j)^2}$

Studied dataset



South of France

Sentinel-2 tiles

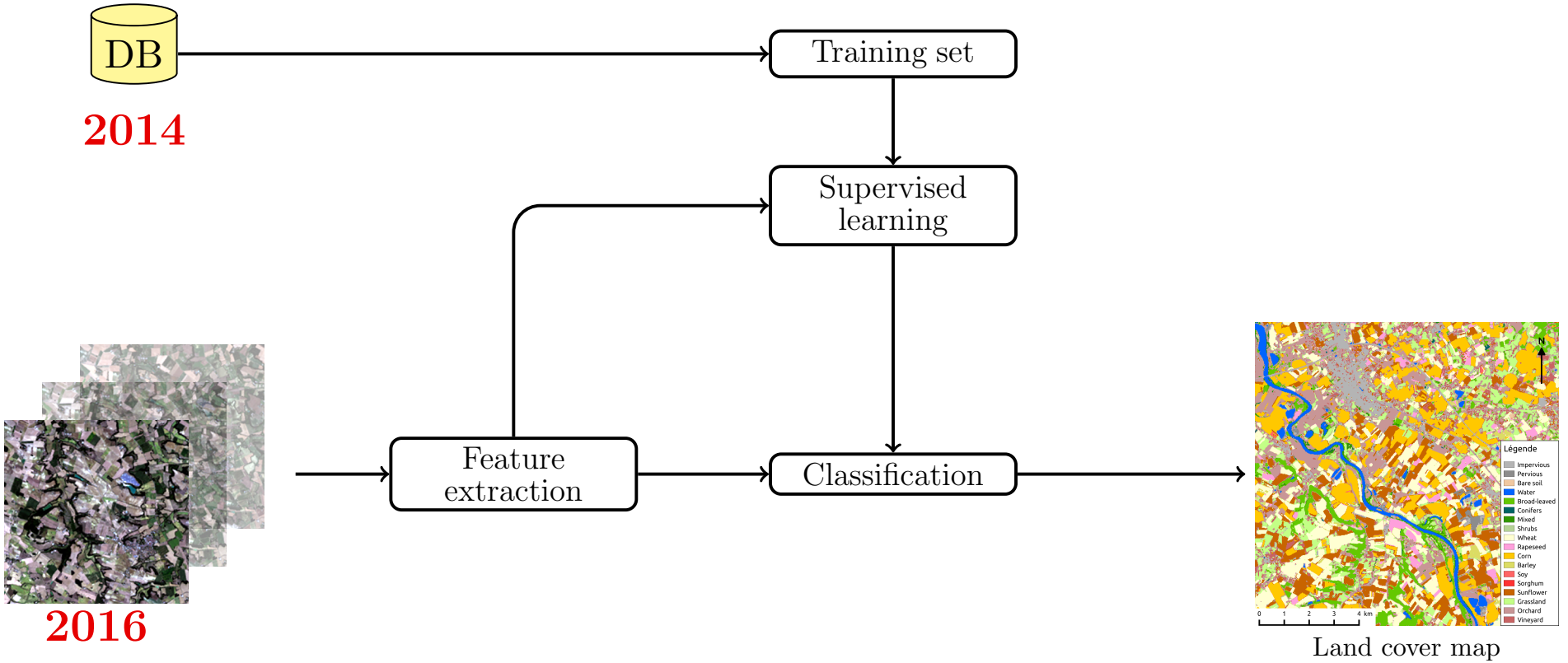
from 30/11/2015 to 15/10/2016

land cover polygons

6 vegetation classes

straw cereals (wheat and barley)
maize, rapeseed, sunflower
vineyard and grassland

Studied dataset



Land cover map

Studied dataset

Sentinel-2 data

Out-of-date reference database

- **training**: farmer's declaration (French Land Parcel Information System) 2014
- satellite data 2016

	no. training samples	no. mislabeled samples	mislabeled %
Straw cereals	500	146	29.2
Maize	500	122	24.4
Rapeseed	500	150	30.0
Sunflower	500	149	29.8
Vineyard	500	0	0.0
Grassland	500	100	20.0

Studied dataset

Sentinel-2 data

Out-of-date reference database

- **training**: farmer's declaration (French Land Parcel Information System) 2014
- satellite data 2016

	no. training samples	no. mislabeled samples	mislabeled %
Straw cereals	500	146	29.2
Maize	500	122	24.4
Rapeseed	500	150	30.0
Sunflower	500	149	29.8
Vineyard	500	0	0.0
Grassland	500	100	20.0

Evaluation

- F-Score values
- parameters of each evaluated methods are optimized

Results

F-Score values

	*k*NN	LOF	ENN	Breiman
Straw cereals	71.3	67.2	77.1	**80.5**
Maize	75.7	60.3	84.3	**89.1**
Rapeseed	77.0	62.8	85.7	**92.2**
Sunflower	74.6	63.8	79.2	**83.7**
Grassland	54.6	53.2	**71.6**	69.0

Results

F-Score values

	*k*NN	LOF	ENN	Breiman
Straw cereals	71.3	67.2	77.1	**80.5**
Maize	75.7	60.3	84.3	**89.1**
Rapeseed	77.0	62.8	85.7	**92.2**
Sunflower	74.6	63.8	79.2	**83.7**
Grassland	54.6	53.2	**71.6**	69.0

- RF outlier method > classical methods
- exception for the grassland class

Results

F-Score values

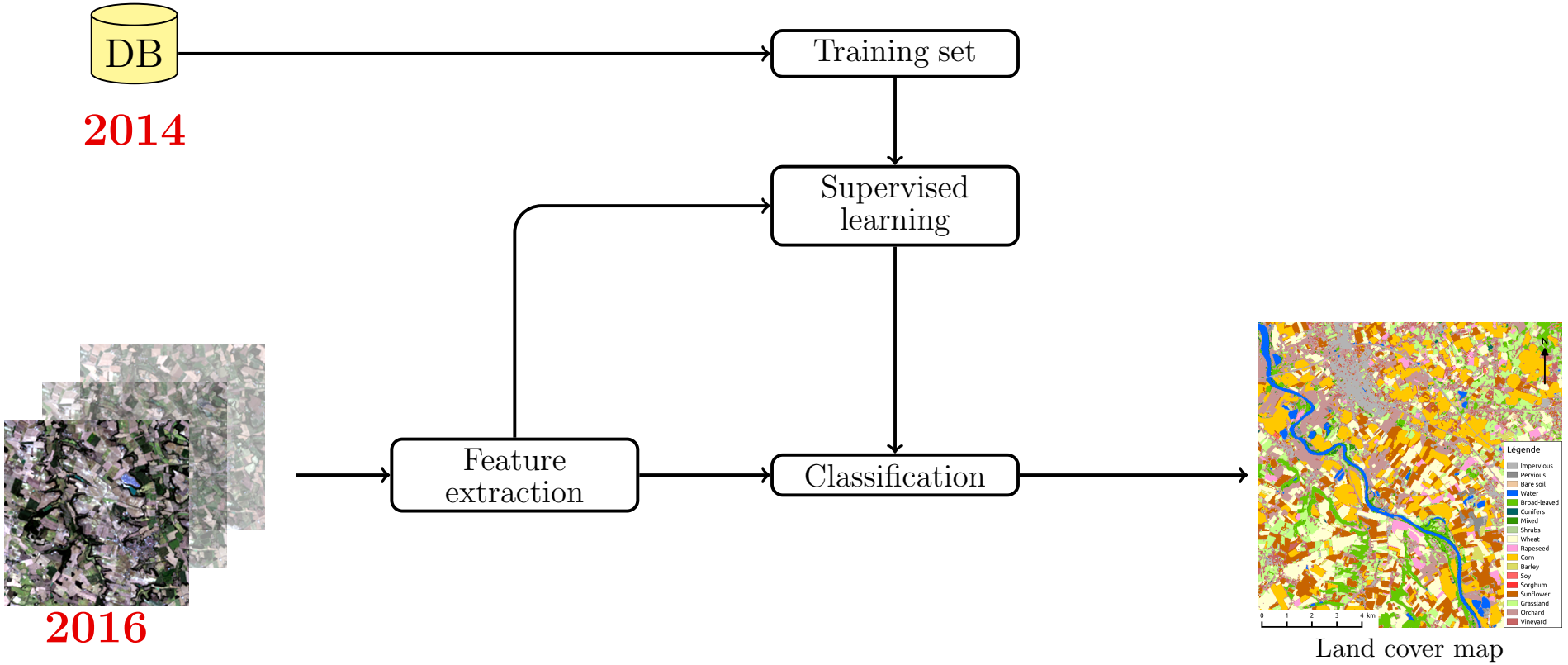
	*k*NN	LOF	ENN	Breiman
Straw cereals	71.3	67.2	77.1	**80.5**
Maize	75.7	60.3	84.3	**89.1**
Rapeseed	77.0	62.8	85.7	**92.2**
Sunflower	74.6	63.8	79.2	**83.7**
Grassland	54.6	53.2	**71.6**	69.0

- RF outlier method > classical methods
- exception for the grassland class

⇒ How to use these outlier detection methods to improve the classification performance?

Problem presentation

How to get an accurate land cover map
when out-of-date training instances are used?



Land cover map

Problem presentation

Reference databases

- **training**: farmer's declaration (French Land Parcel Information System) 2014
- **testing**: field campaigns 2016

	no. validation instances (2016)	no. training instances (2014)	mislabeled %
Straw cereals	197,992		29.2
Maize	125,573		24.4
Rapeseed	43,384	500	30.0
Sunflower	115,618		29.8
Vineyard	11,973		0.0
Grassland	72,200		20.0

Problem presentation

Reference databases

- **training**: farmer's declaration (French Land Parcel Information System) 2014
- **testing**: field campaigns 2016

	no. validation instances (2016)	no. training instances (2014)	mislabeled %
Straw cereals	197,992		29.2
Maize	125,573		24.4
Rapeseed	43,384	500	30.0
Sunflower	115,618		29.8
Vineyard	11,973		0.0
Grassland	72,200		20.0

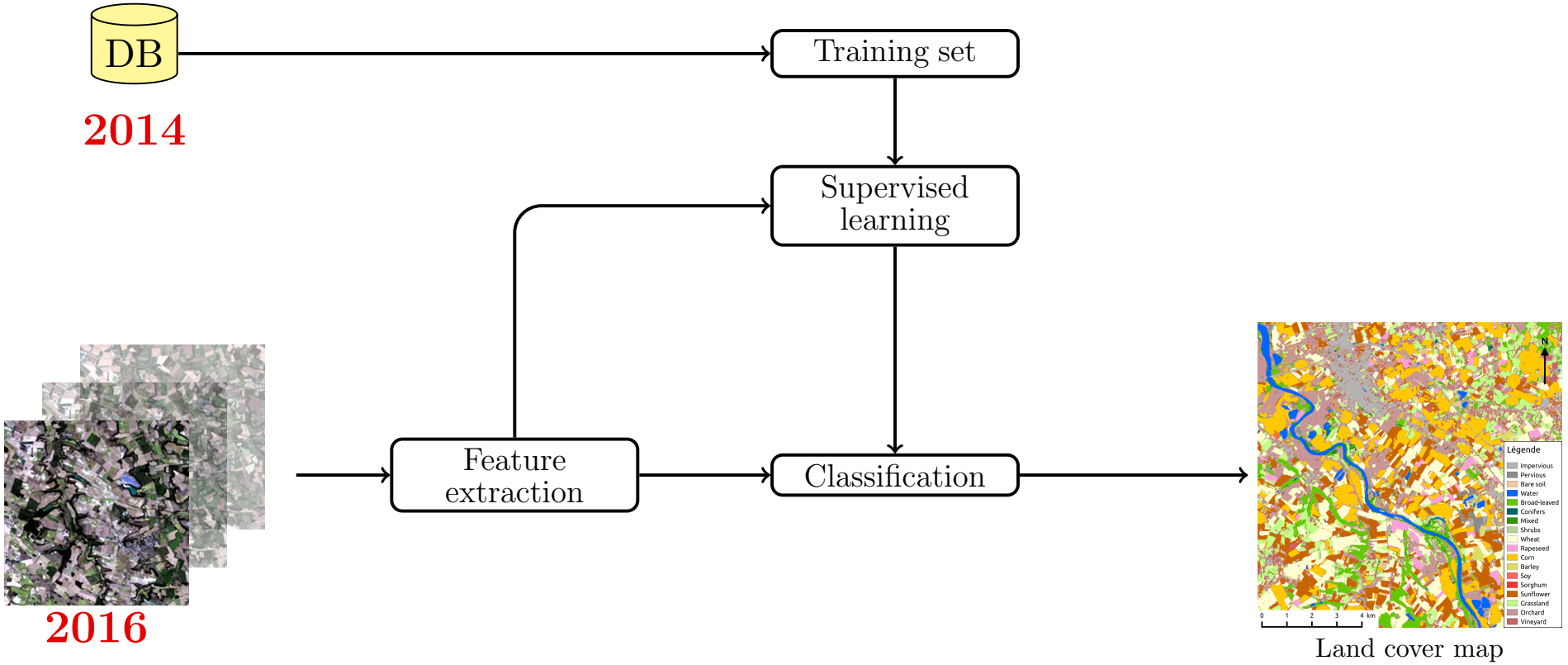
Influence of mislabeled data on classification performance

- 2016 clean training dataset versus 2014 mislabeled training dataset
- Overall Accuracy (OA) and F-Score values

	2016	2014
Straw cereals	95.5	79.5
Maize	96.4	84.3
Rapeseed	90.7	40.0
Sunflower	95.7	79.6
Vineyard	84.4	78.8
Grassland	89.8	79.2
OA	94.4 ± 0.6	78.2 ± 0.2

Filtering mislabeled training data

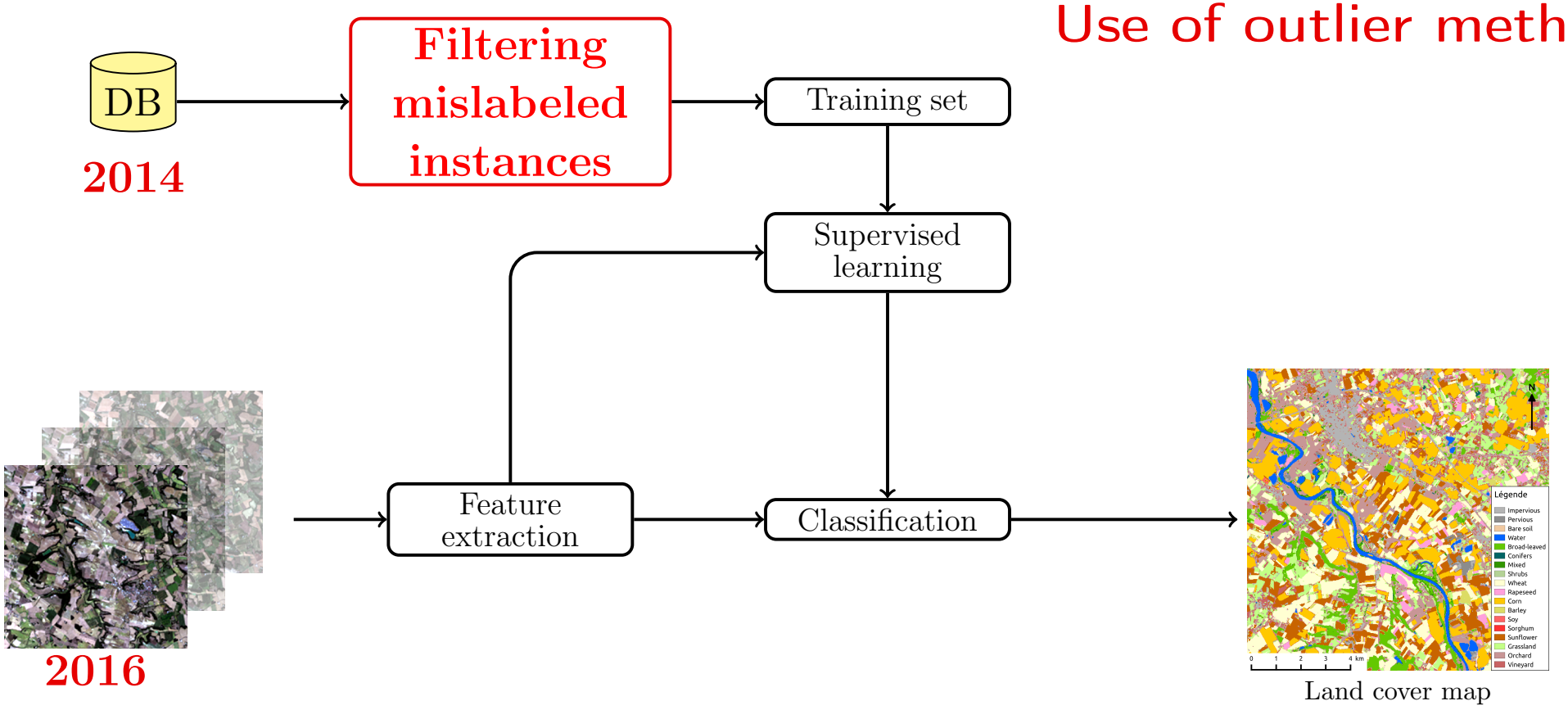
How to get an accurate land cover map
when out-of-date training instances are used?



Filtering mislabeled training data

How to get an accurate land cover map
when out-of-date training instances are used?

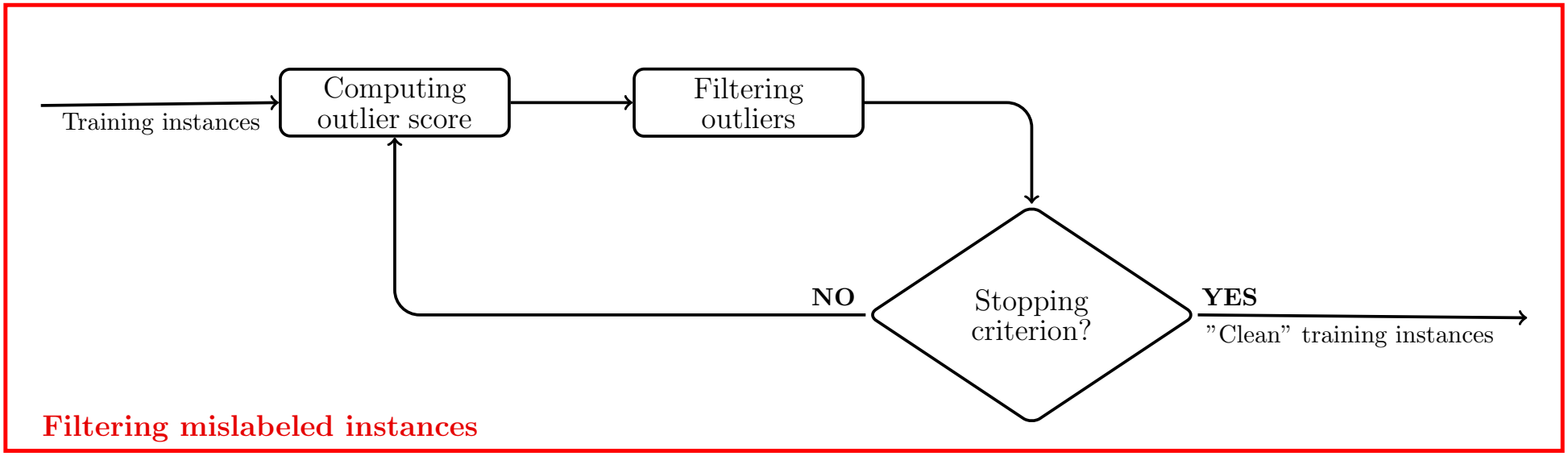
Use of outlier methods



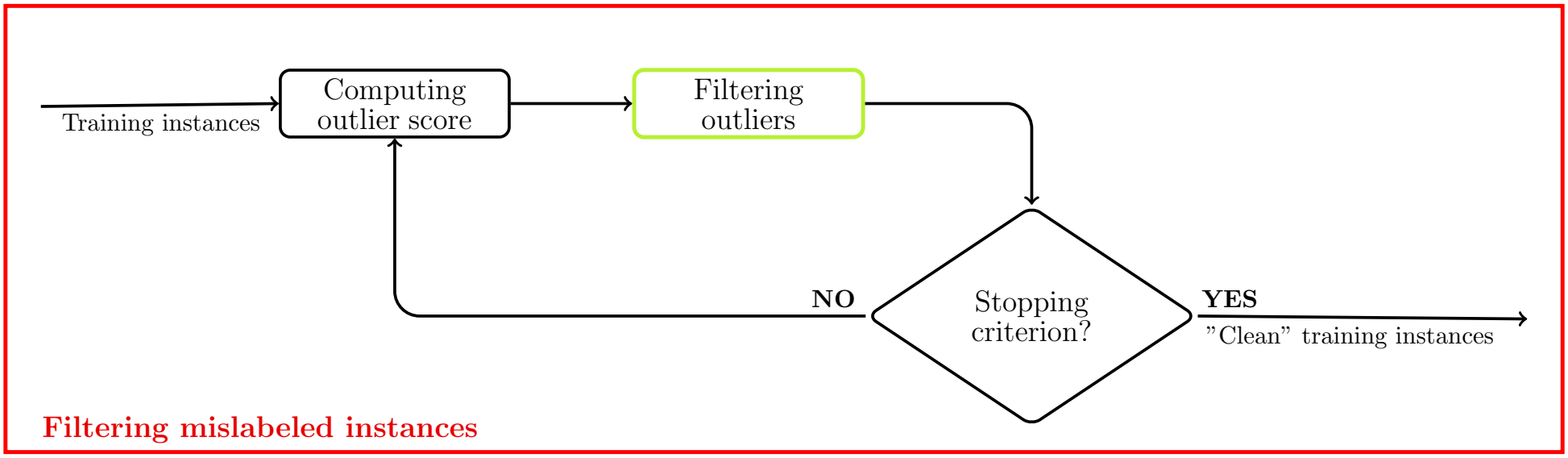
Land cover map

Filtering mislabeled training data

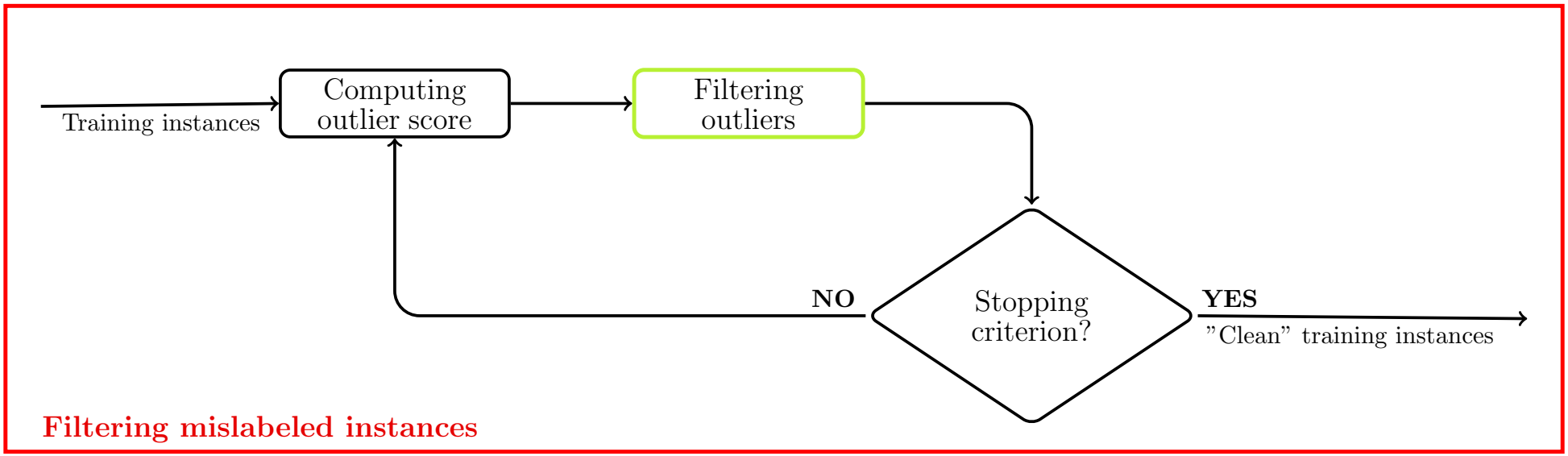
The proposed framework



1. Filtering outliers



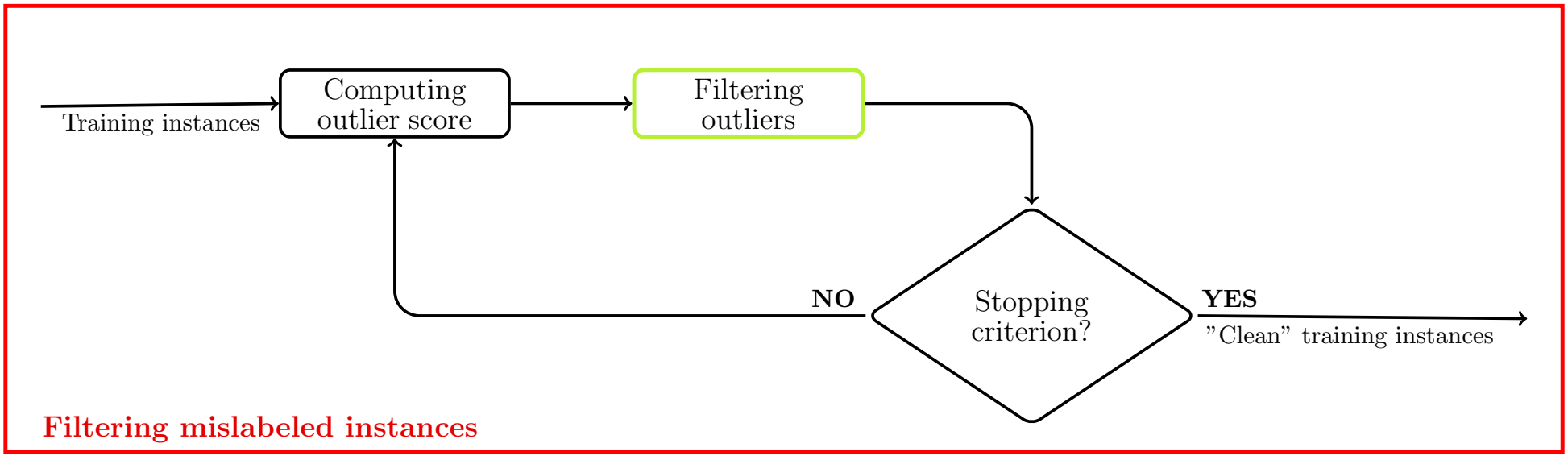
1. Filtering outliers



Two strategies

- Global filtering
  - the n instances with the highest outlier scores are removed

1. Filtering outliers



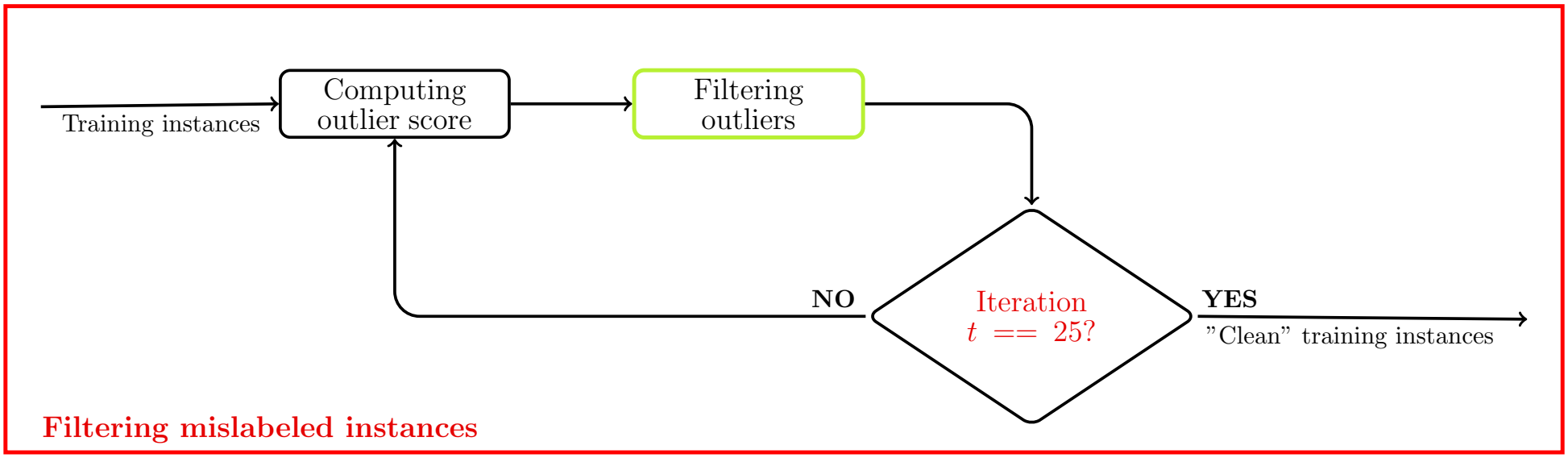
Two strategies

- Global filtering
  - the n instances with the highest outlier scores are removed
- Class-based filtering
  - the instances that do not fulfilled the 3σ rule per class are removed
  - $OS(i) \geq \overline{OS} + 3\sigma_{OS}$

1. Filtering outliers

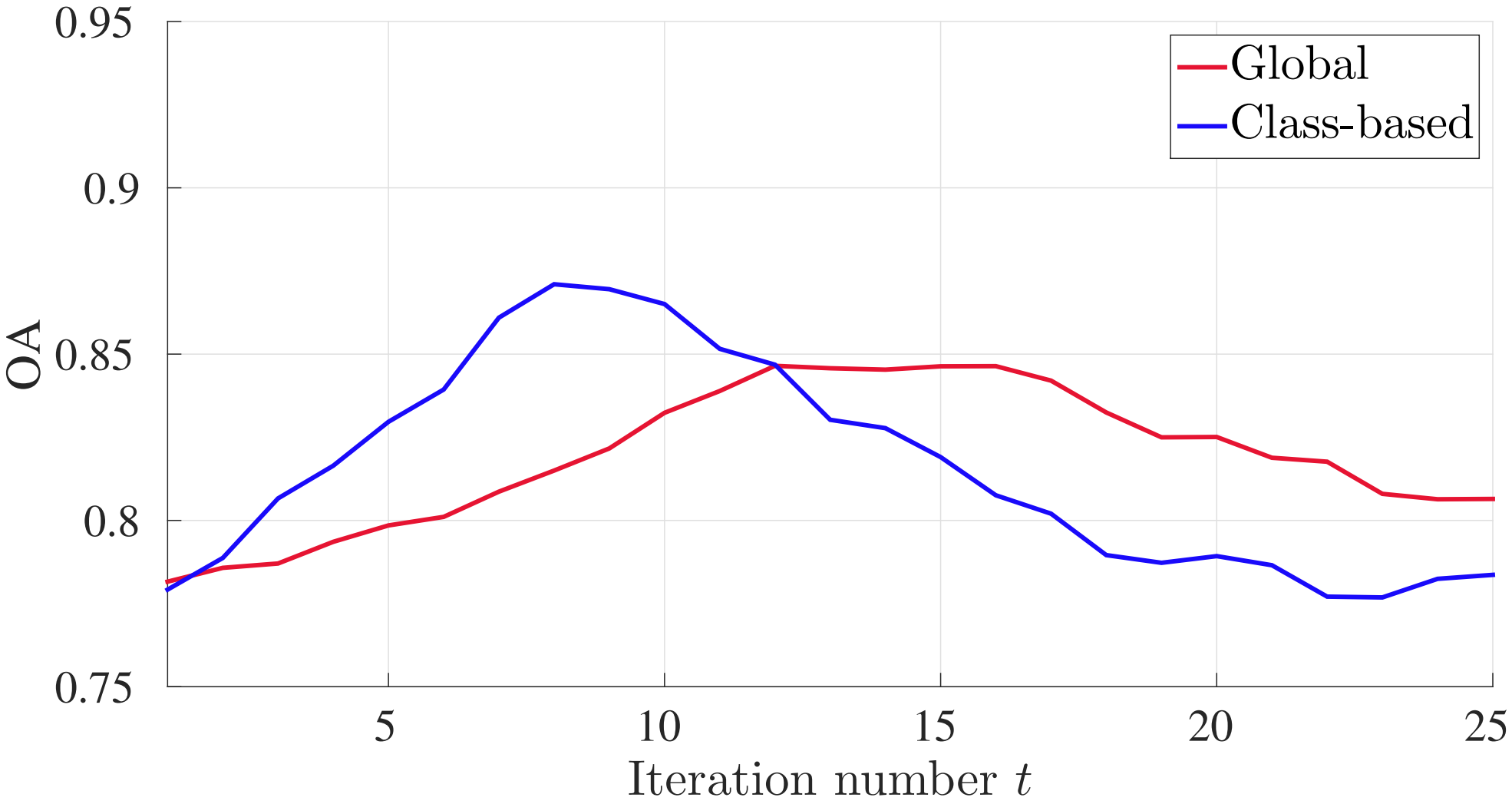
Evaluation of the proposed iterative procedures

- evaluation of OA obtained at each iteration with the training dataset filter by the proposed procedures



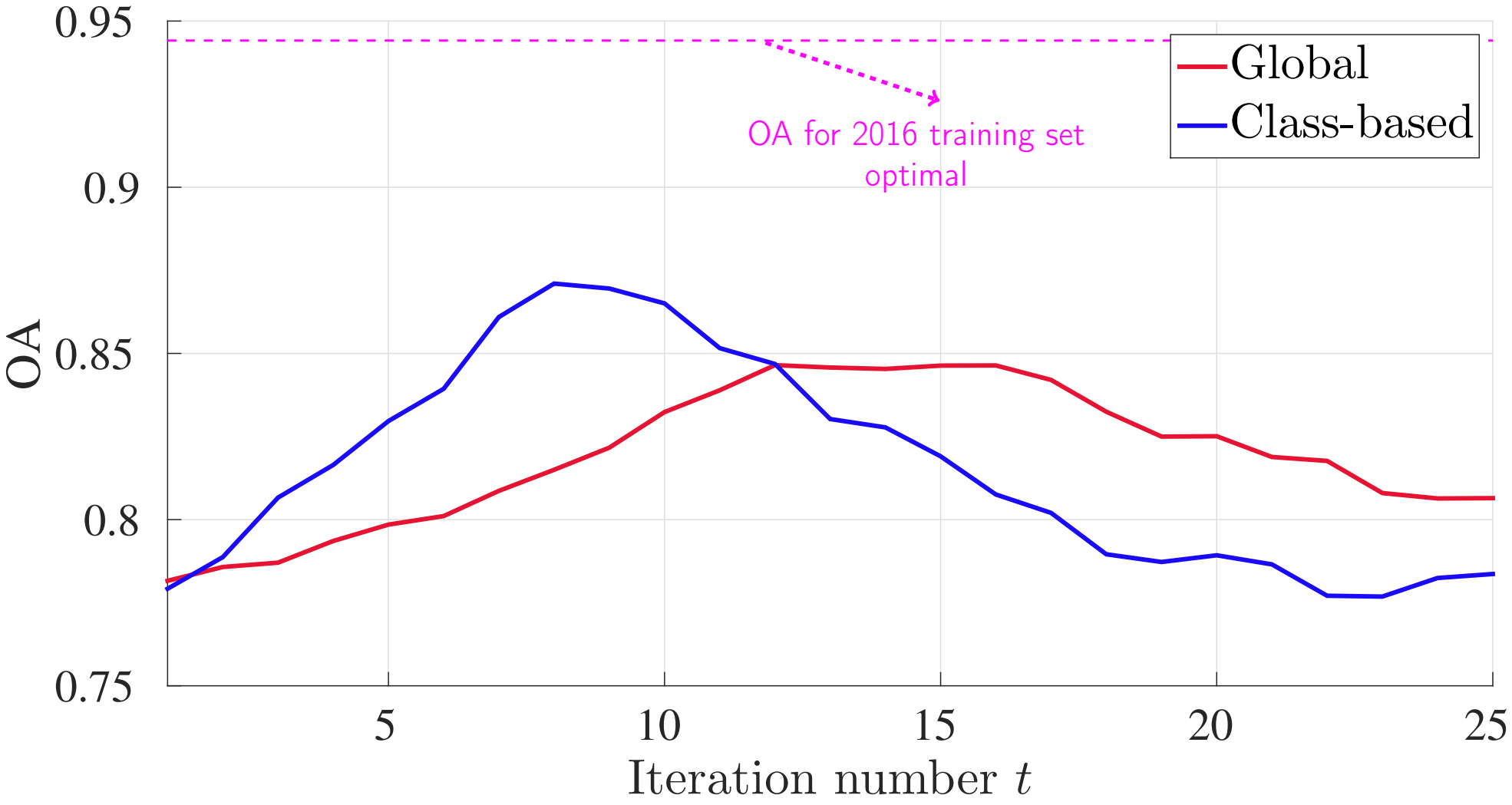
1. Filtering outliers

Evaluation of the proposed iterative procedures



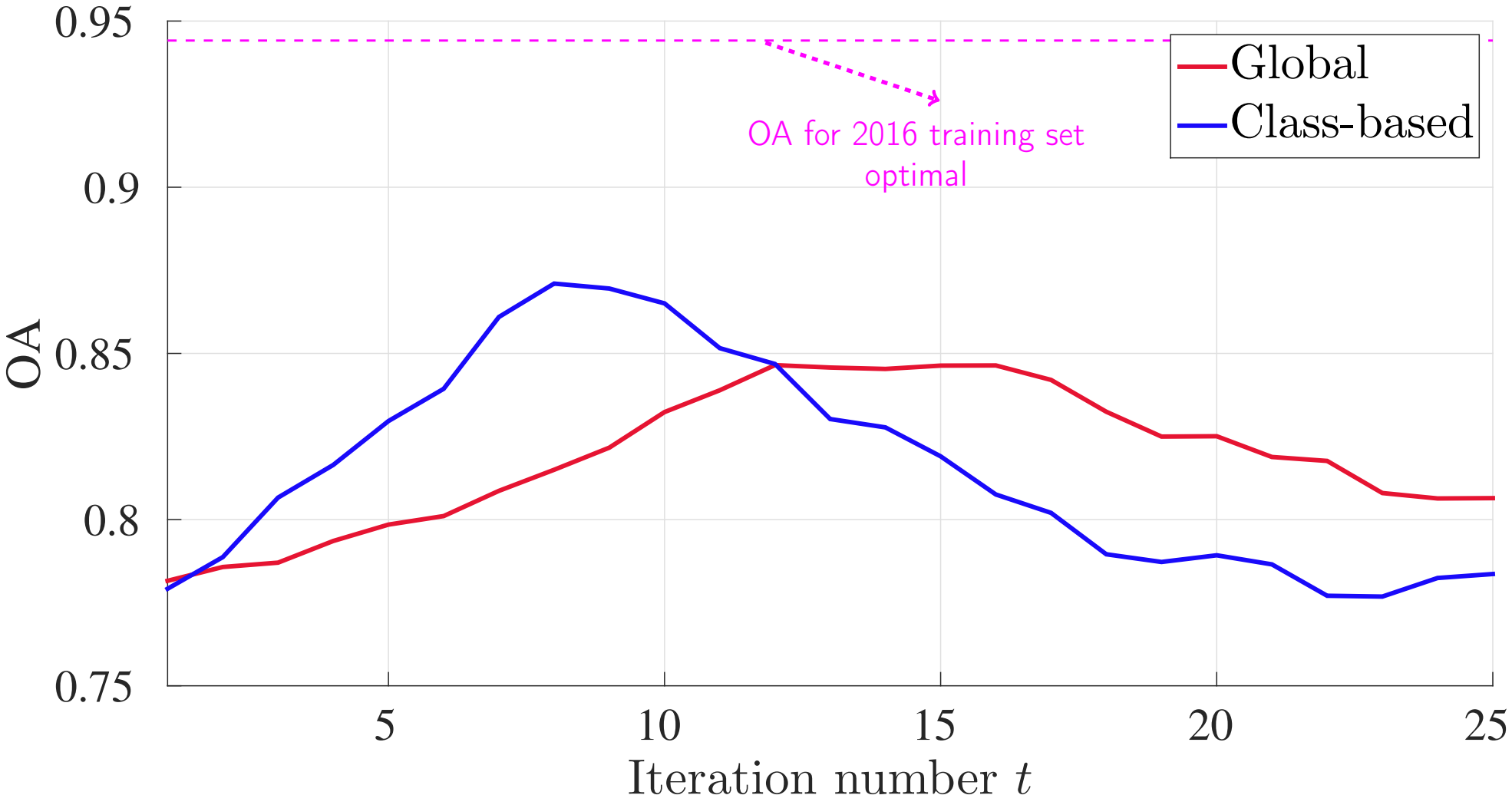
1. Filtering outliers

Evaluation of the proposed iterative procedures



1. Filtering outliers

Evaluation of the proposed iterative procedures

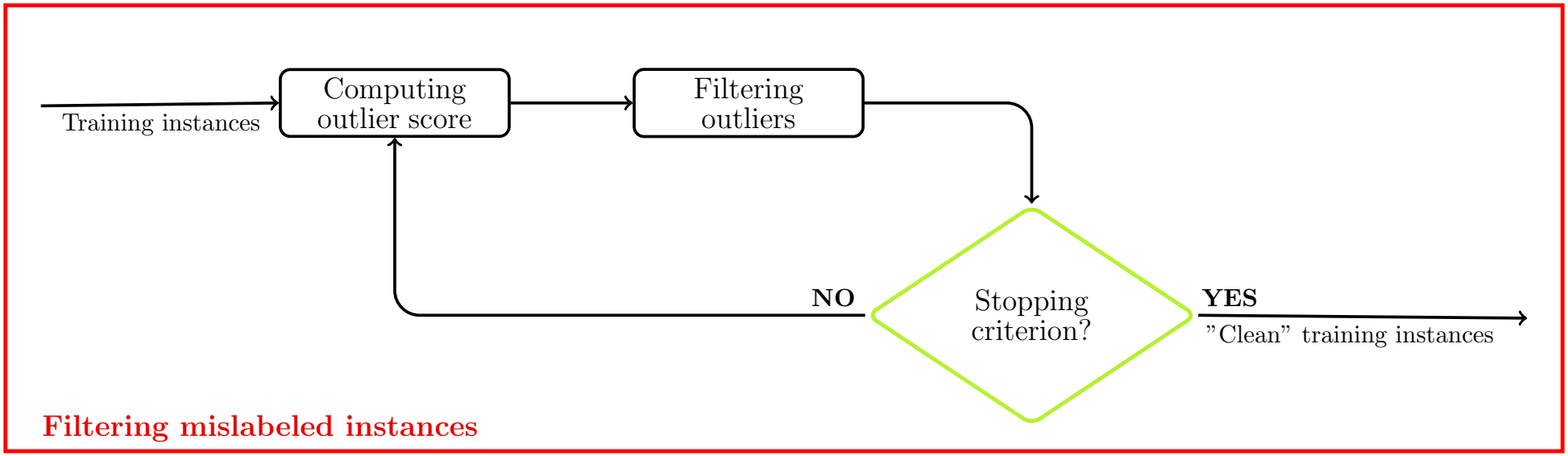


When the procedure should be stopped?

2. Stopping criterion

Challenges

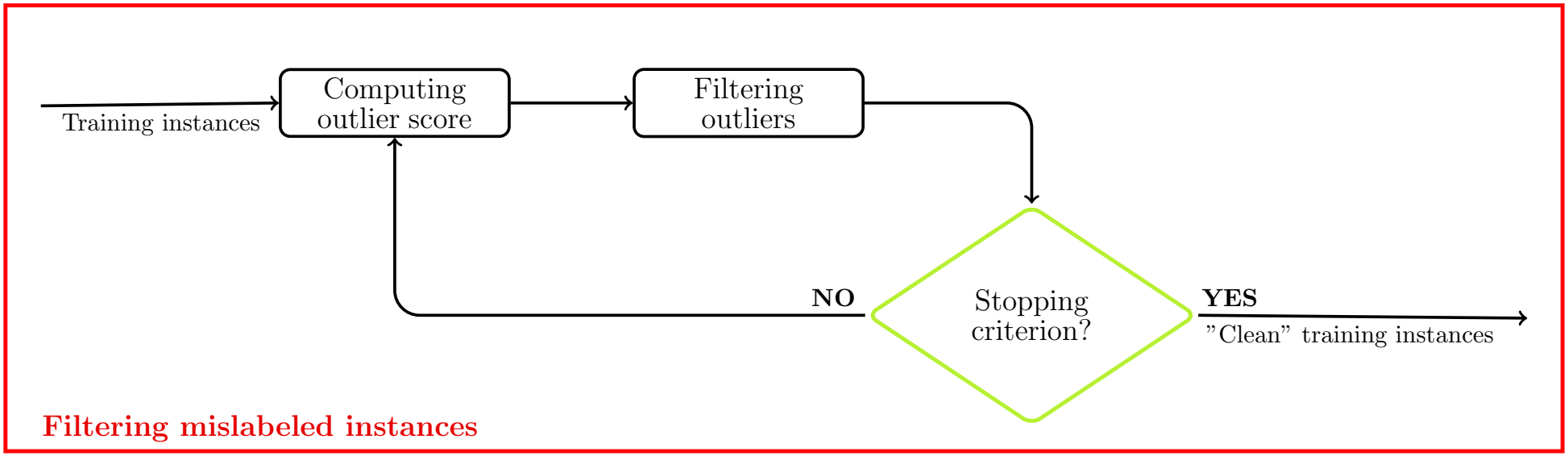
- the stopping criterion needs to be independent from 2016 validation dataset



2. Stopping criterion

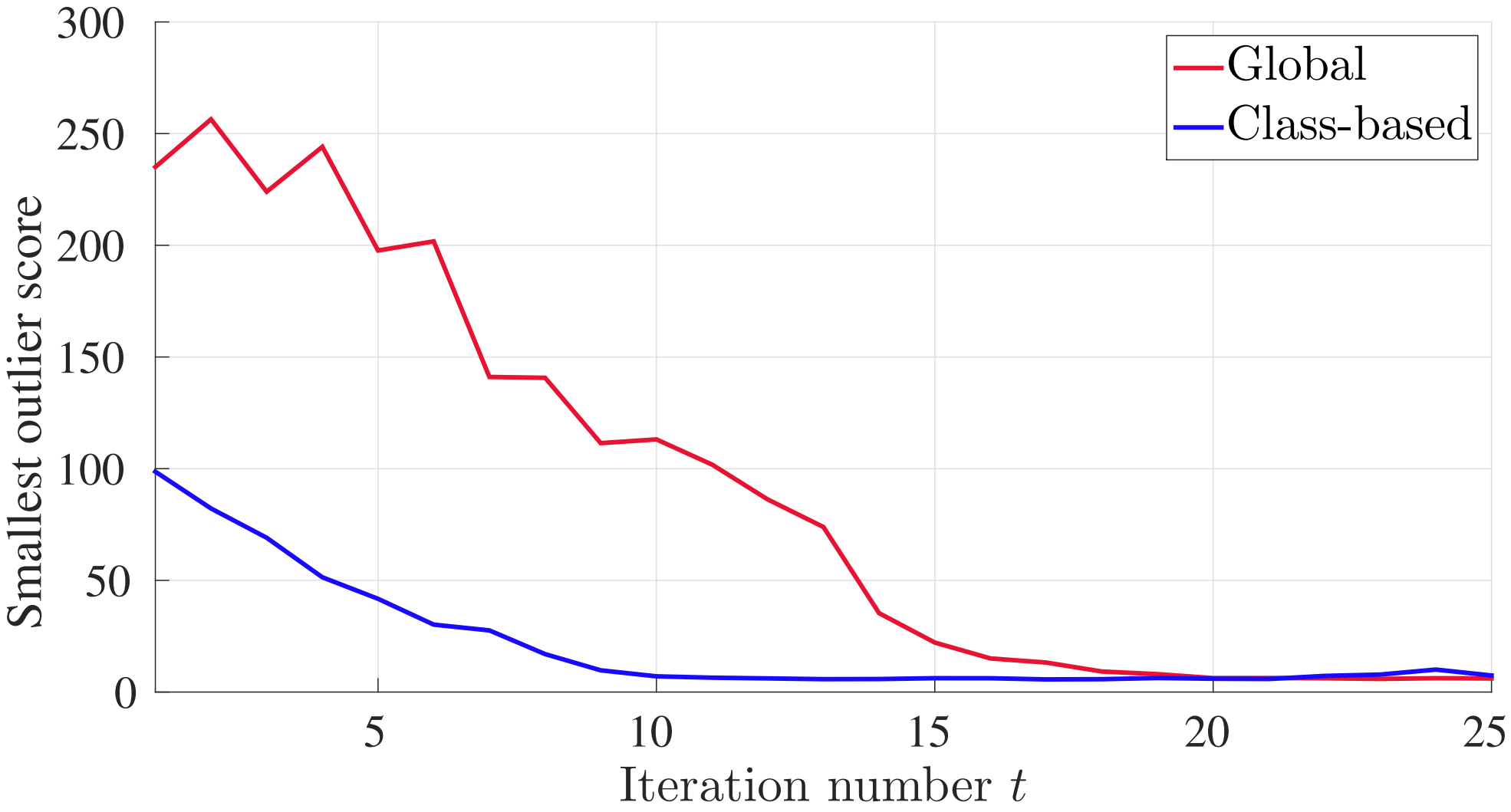
Challenges

- the stopping criterion needs to be independent from 2016 validation dataset



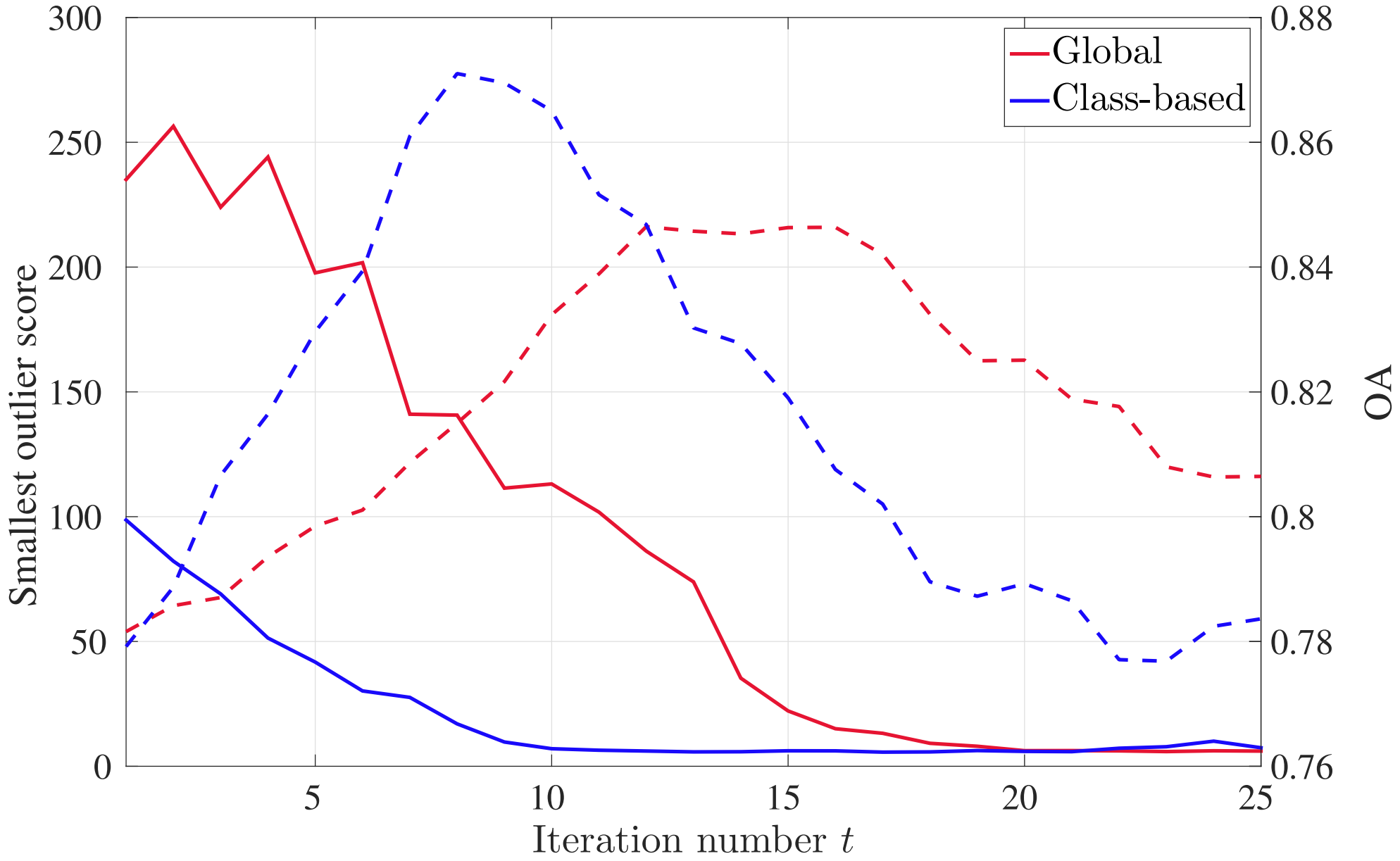
Use of the smallest outlier score among the removed instances at each iteration

2. Stopping criterion



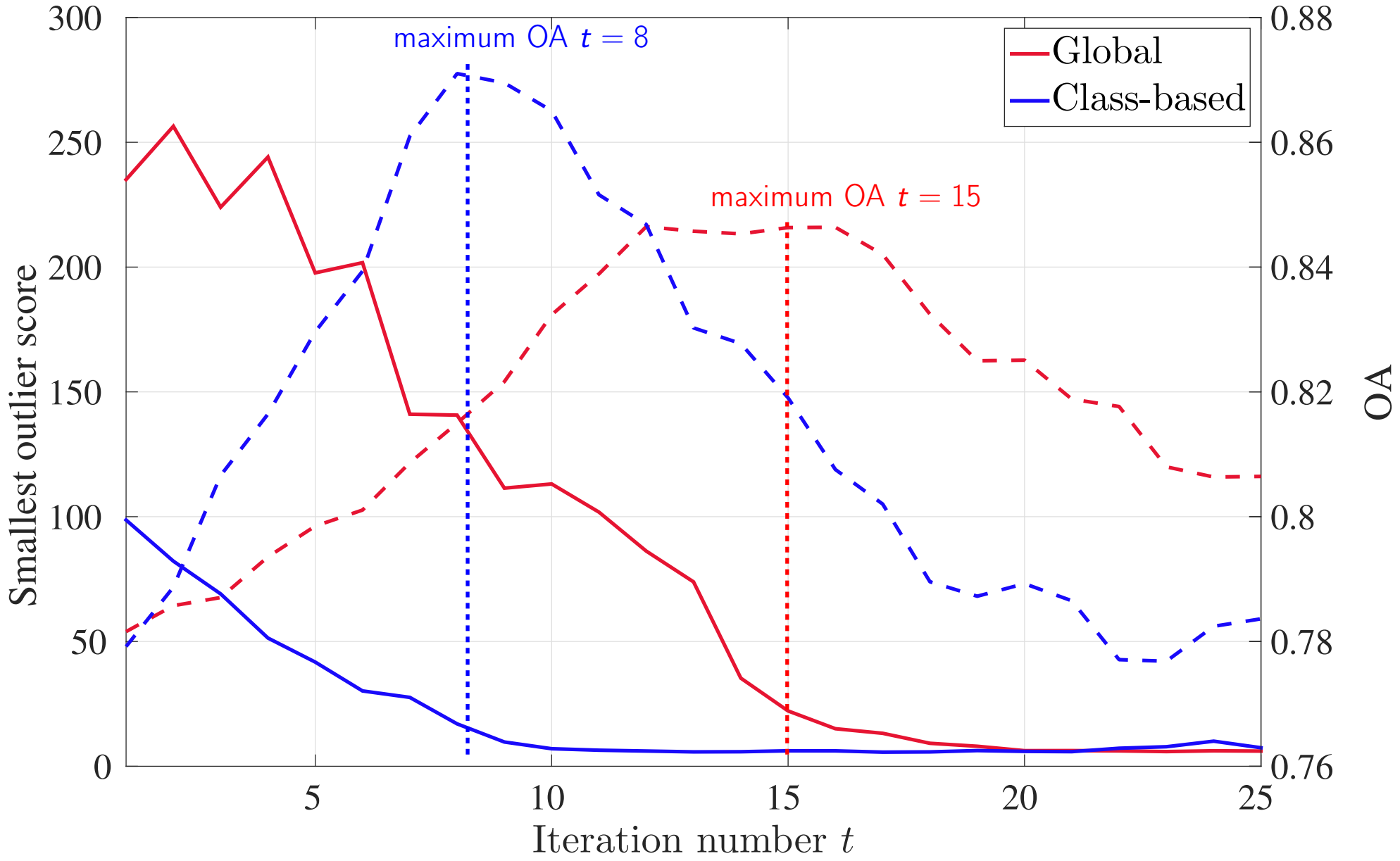
2. Stopping criterion

- Correlation between the smallest outlier scores and OA values



2. Stopping criterion

- Correlation between the smallest outlier scores and OA values



Summary

Percentage of mislabeled data in the original and filtered training datasets

	no. validation instances (2016)	**2014 original** $t = 0$	**2014 global** $t = 15$	**2014 class-based** $t = 8$	**2014 ENN** $k = 21$
Straw cereals	197,992	29.2	2.0	3.8	1.5
Maize	125,573	24.4	0.6	0.3	1.5
Rapeseed	43,384	30.0	0.6	0.0	0.0
Sunflower	115,618	29.8	5.1	4.7	0.1
Vineyard	11,973	0.0	0.0	0.0	0.0
Grassland	72,200	20.0	11.6	2.8	6.7

Summary

Percentage of mislabeled data in the original and filtered training datasets

	no. validation instances (2016)	2014 original $t = 0$	2014 global $t = 15$	2014 class-based $t = 8$	2014 ENN $k = 21$
Straw cereals	197,992	29.2	2.0	3.8	1.5
Maize	125,573	24.4	0.6	0.3	1.5
Rapeseed	43,384	30.0	0.6	0.0	0.0
Sunflower	115,618	29.8	5.1	4.7	0.1
Vineyard	11,973	0.0	0.0	0.0	0.0
Grassland	72,200	20.0	11.6	2.8	6.7

Classification performance

	2016	2014 original $t = 0$	2014 global $t = 15$	2014 class-based $t = 8$	2014 ENN $k = 21$
Straw cereals	95.5	79.5	85.4	90.1	**88.5**
Maize	96.6	84.3	**91.0**	88.7	90.7
Rapeseed	91.4	40.0	**79.8**	78.1	79.6
Sunflower	95.8	79.6	**91.2**	90.5	90.7
Vineyard	84.2	78.8	**60.5**	57.1	58.3
Grassland	89.5	79.2	73.4	**79.2**	77.1
OA	94.4	78.2	84.6	**86.5**	86.0

Conclusions

Label noise is one of the most key challenge in remote sensing

- it influences the classification performance (and not only)
- detection of mislabeled training data is possible
- filtered the mislabeled training data improves the classification performance, but there is still a gap...

Conclusions

Label noise is one of the most key challenge in remote sensing

- it influences the classification performance (and not only)
- detection of mislabeled training data is possible
- filtered the mislabeled training data improves the classification performance, but there is still a gap...

Outlook

- increasing the number of instances
- analysing the detected mislabeled instances
  - correct them
  - active learning
  - semi supervised learning
- use the proposed methods for change detection

Conclusions

Label noise is one of the most key challenge in remote sensing

- it influences the classification performance (and not only)
- detection of mislabeled training data is possible
- filtered the mislabeled training data improves the classification performance, but there is still a gap...

Outlook

- increasing the number of instances
- analysing the detected mislabeled instances
  - correct them
  - active learning
  - semi supervised learning
- use the proposed methods for change detection

Thank you for your attention

Questions?

Influence of the input feature vectors

For real datasets:

Spectral bands

Influence of the input feature vectors

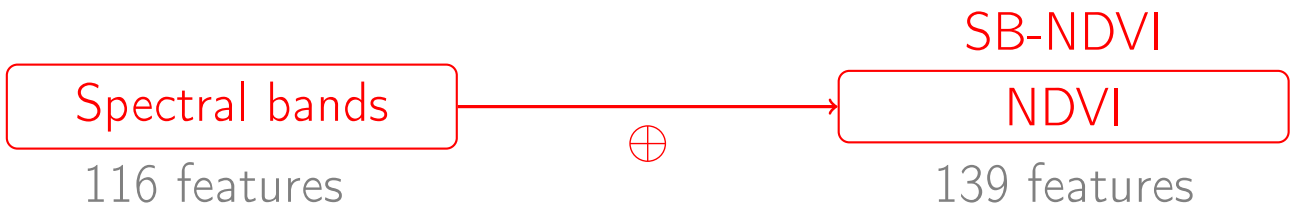
For real datasets:

Spectral bands

NDVI

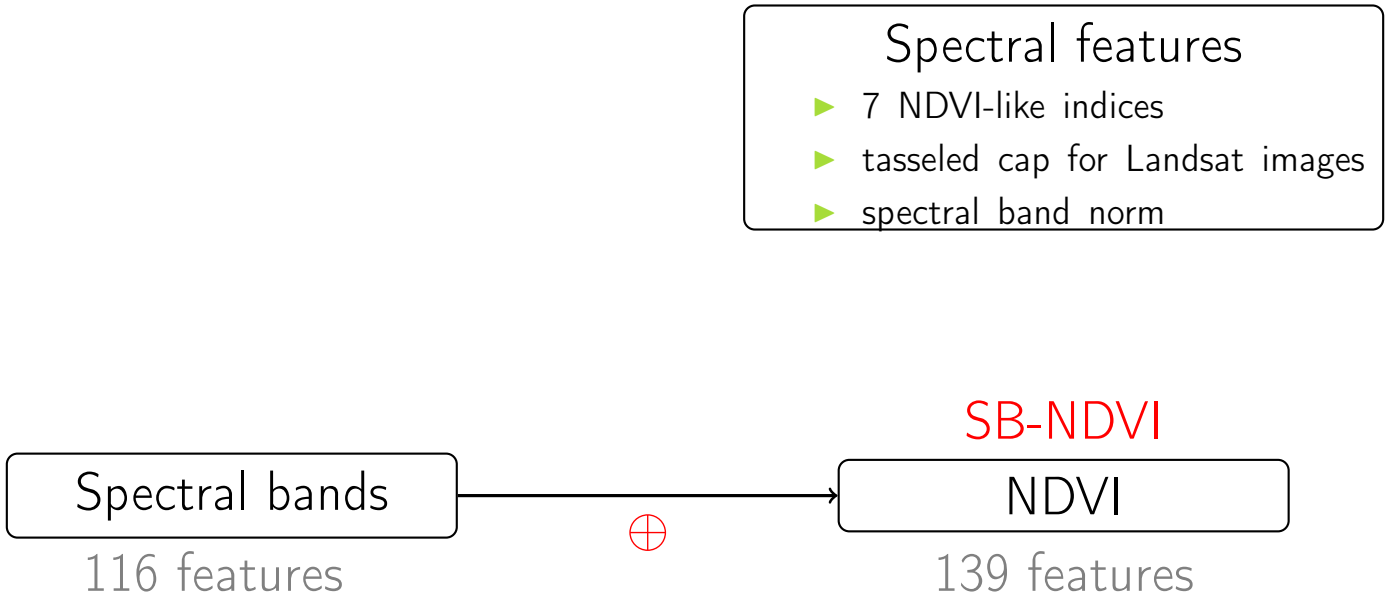
Influence of the input feature vectors

For real datasets:



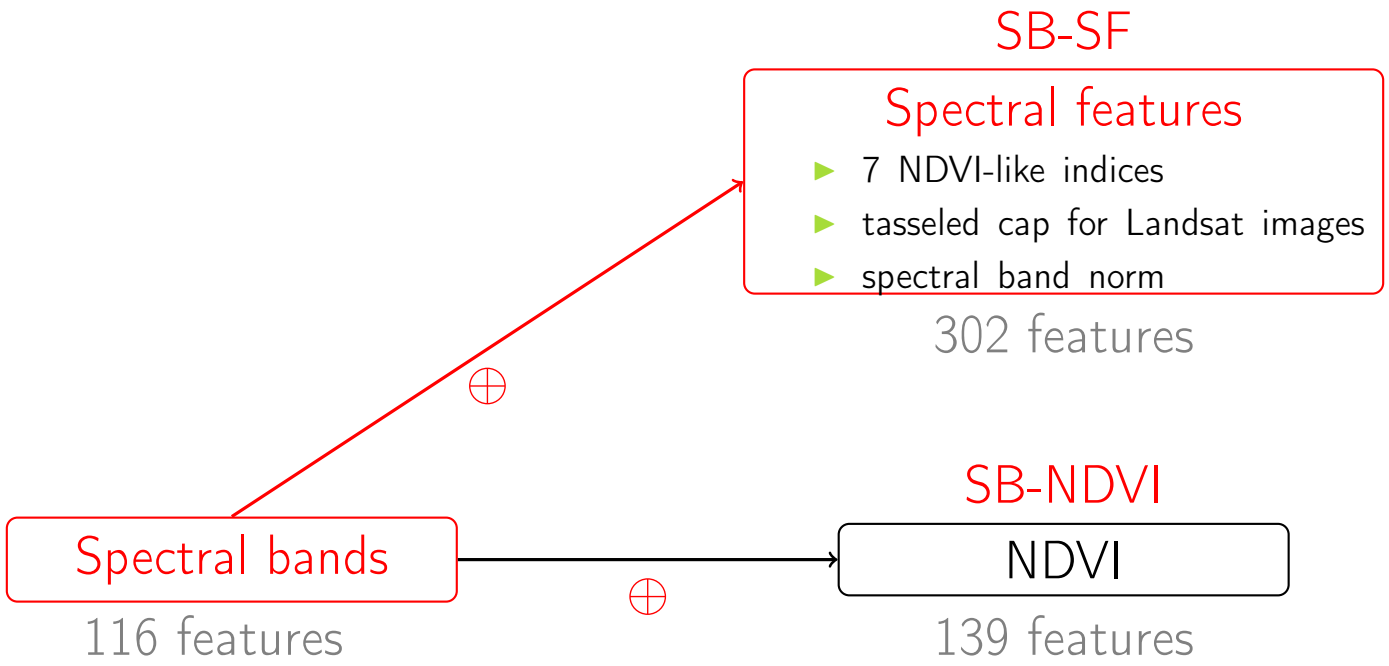
Influence of the input feature vectors

For real datasets:



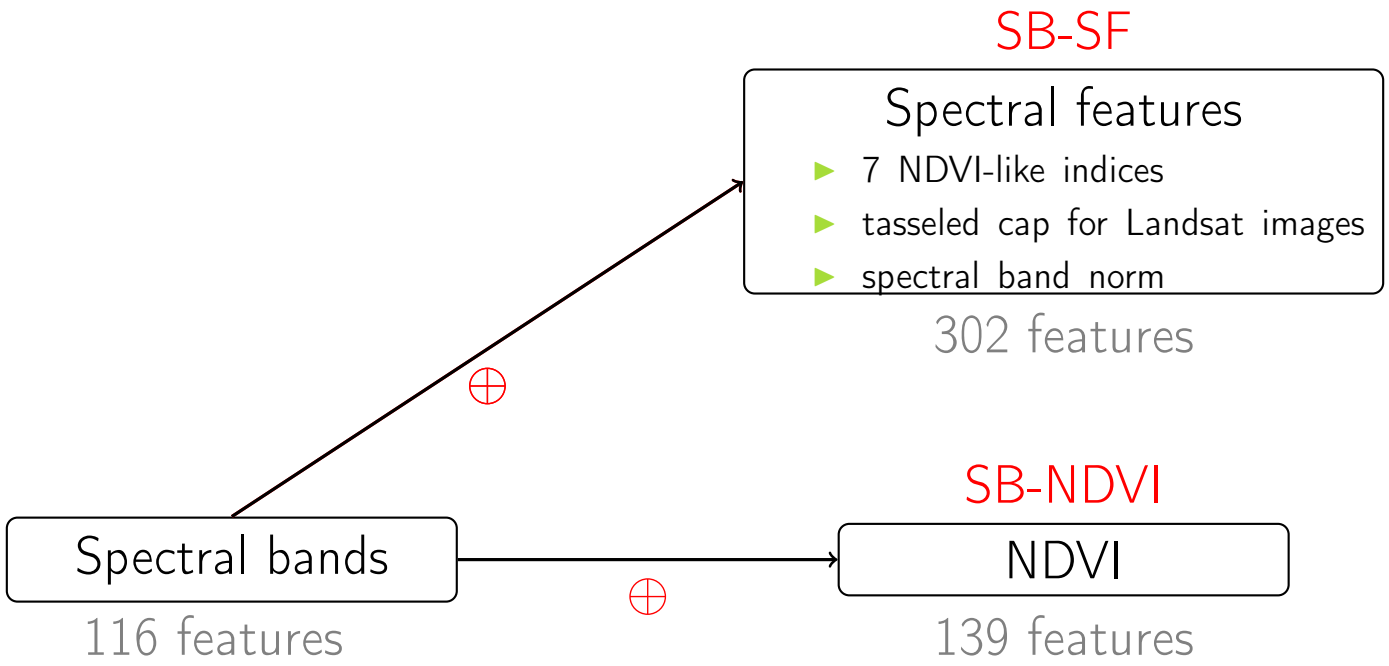
Influence of the input feature vectors

For real datasets:

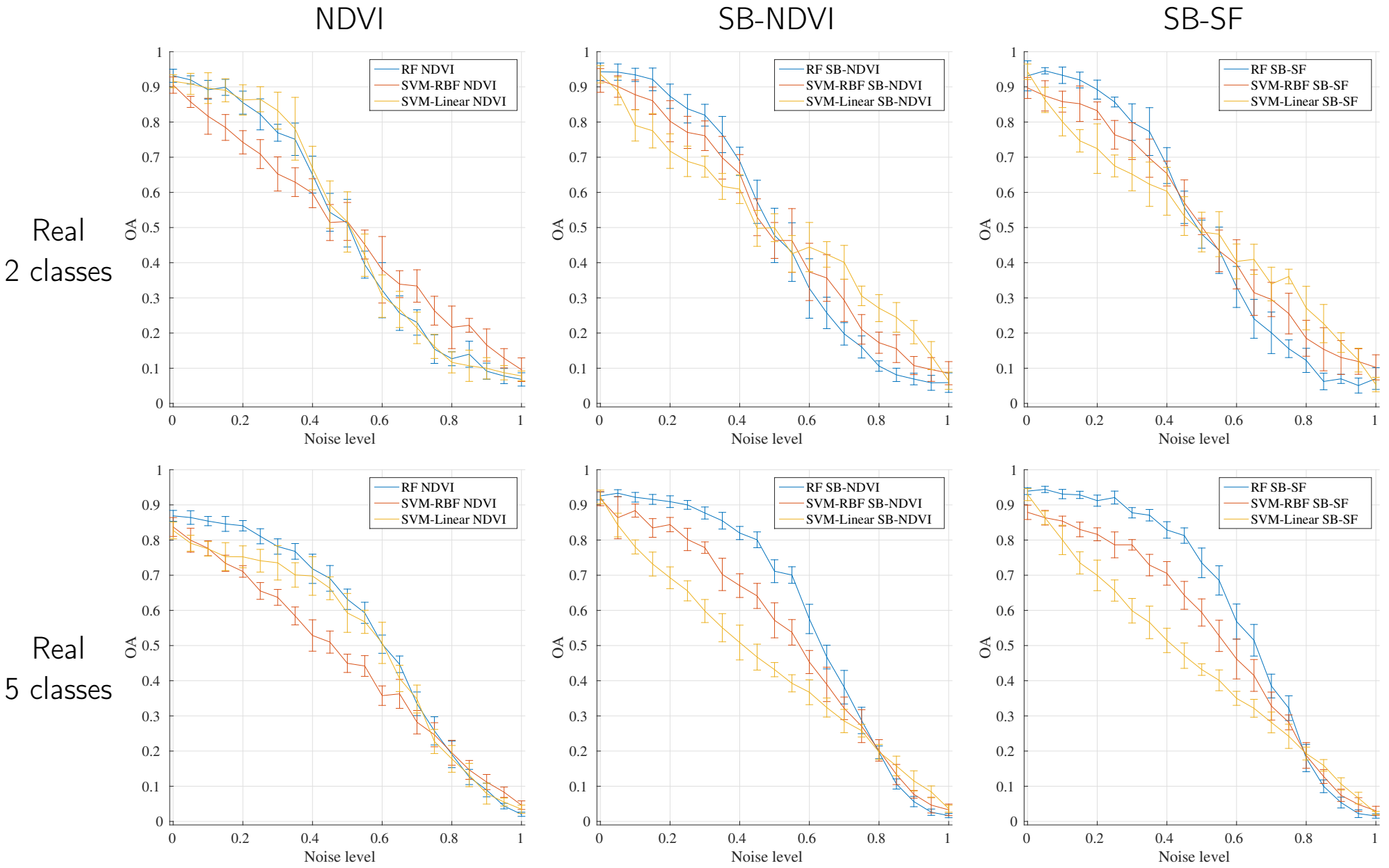


Influence of the input feature vectors

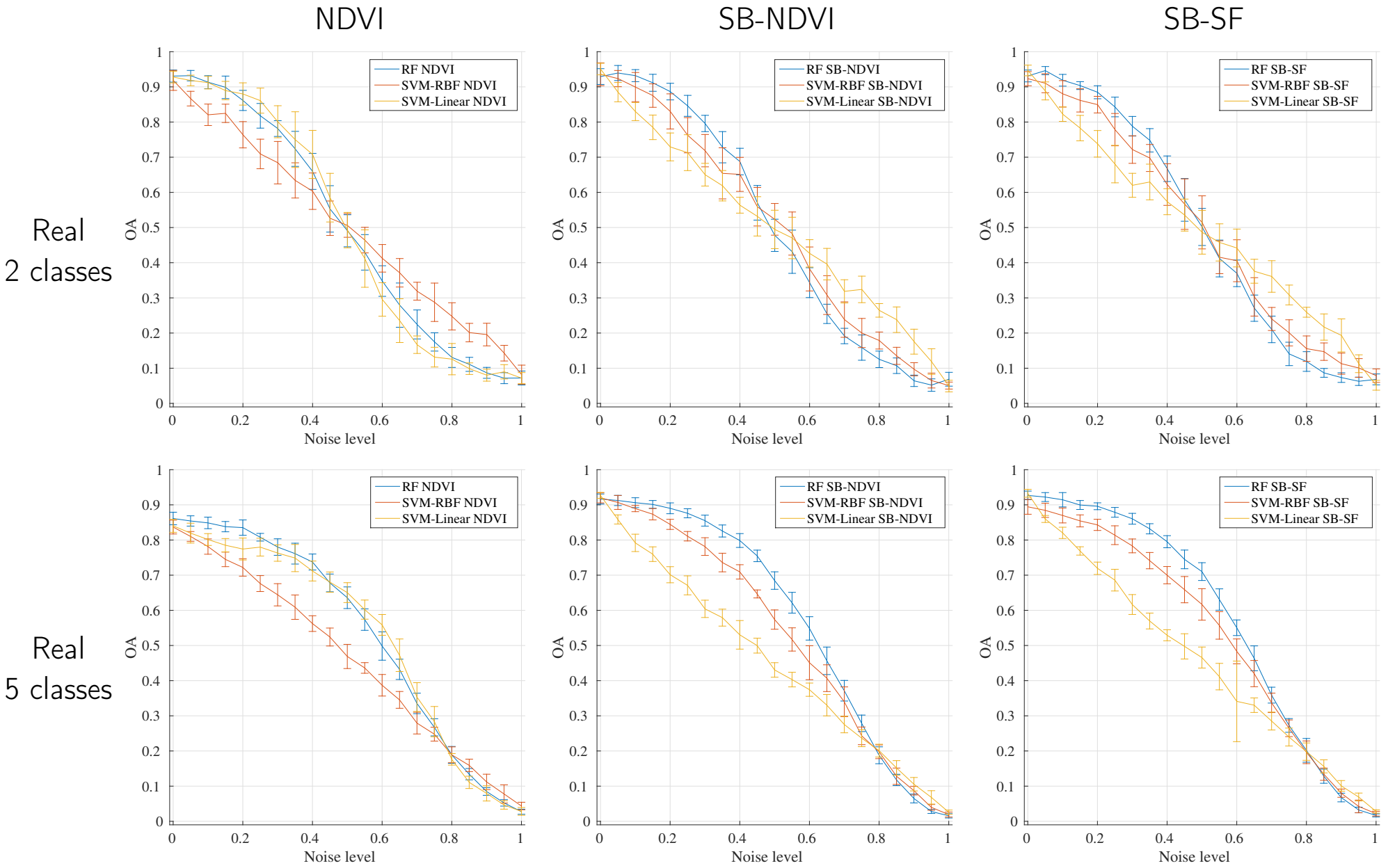
For real datasets:



Influence of the input feature vectors



Influence of the number of training instances



License