STIC-AmSud 2nd Meeting

Curitiba, Oct. 23, 2014

Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes

Prof. Laurent Heutte

Laurent.Heutte@univ-rouen.fr

http://www.litislab.eu/Members/lheutte

University of Rouen, France

- Located in the north of Paris (100 km)
- 33000 students
- 7 Faculties (research/teaching)
  - ✓ Medicine
  - ✓ Sciences
  - ✓ Literacy
  - ✓ Law
  - ✓ Technology
  - ✓ Economic sciences
  - ✓ Psychology

Faculty of Sciences and Techniques

- 3300 students
- 66 diplomas
- 400 professors and researchers
- 200 administrative staff
- Faculty divided into:
  - ✓ 7 departments (teaching): Computer Science, Computer Engineering, Physics, Biology, Mathematics,…
  - ✓ 14 laboratories (research): LITIS, CORIA, IRCOF, … some may associated with CNRS, INRIA, INSERM,…

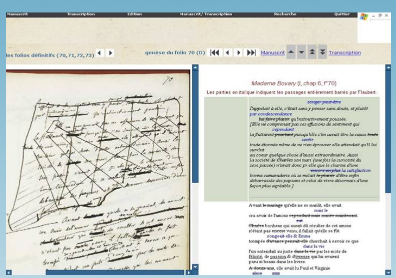
LITIS Lab. (http://www.litislab.eu)

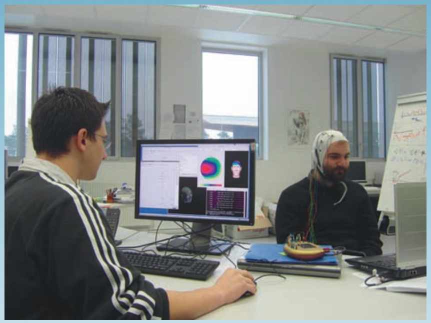
- Laboratory of Computer Science, Information Processing and Systems
- Depending on 3 organizations located in Upper Normandy: University of Rouen, University of Le Havre, INSA Rouen
- Scope: Sciences and Technology of Information and Communications
  - ✓ All formal and practical aspects of *information processing*
- 90 faculty members (whose 31 Prof, 5 Assoc. Prof., 54 Ass. Prof.)
- 7 research teams
- 80 PhD students and post-doc
- Head of the Lab: Prof. T. Paquet

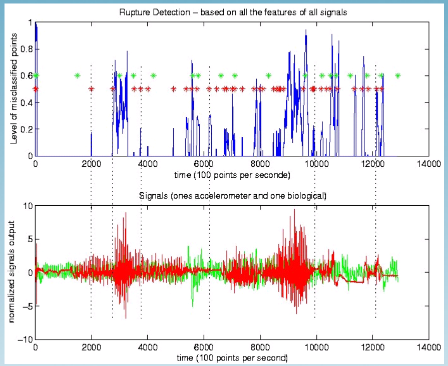
Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes

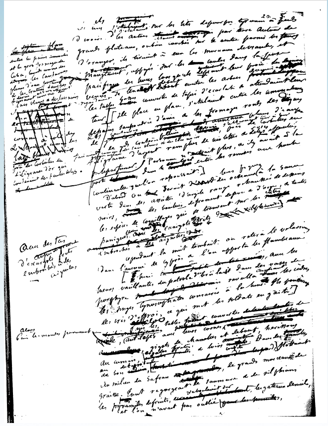
Document and Learning Team

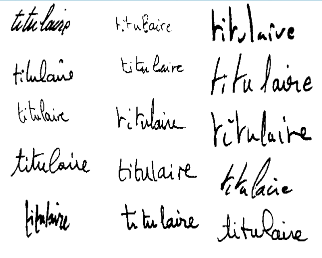
- Head: Prof. L Heutte
- 16,5 staff members – 7 PR, 9.5 Ass. Prof.
- 4 post-doc and enginers
- 16 PhD students

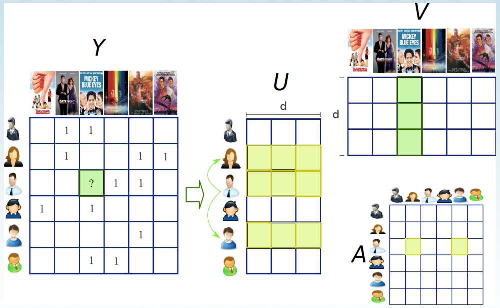


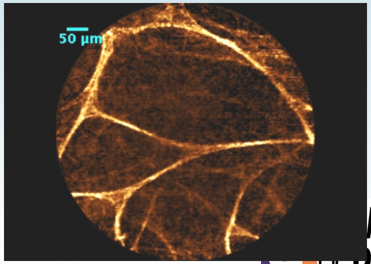


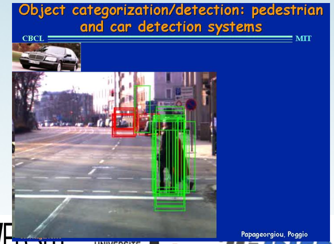












Scientific Issues

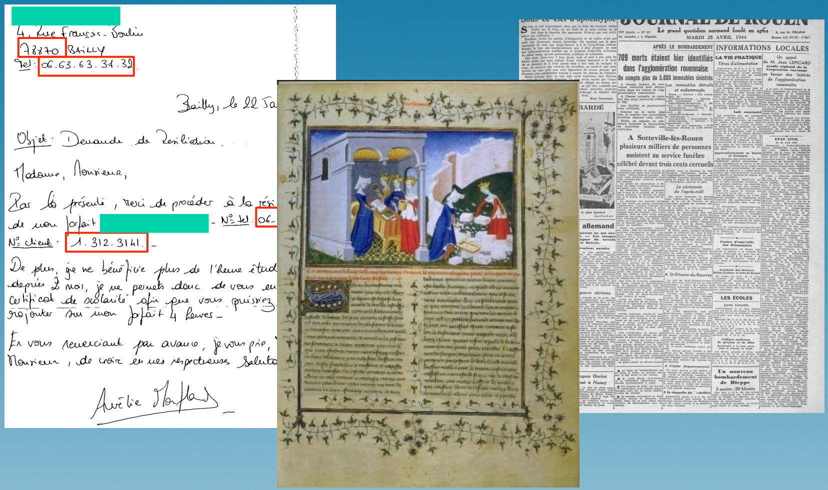
- Machine Learning and Pattern Recognition

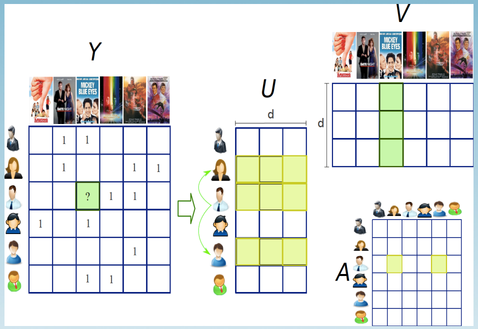
- Joint learning of representations and decisions
  - ✓ Dictionary learning and variable selection, deep learning
  - ✓ Kernel learning (SVM, Kernel PCA, SimpleMKL, regularization path)
  - ✓ Graphs and learning (isomorphism, classification,…)
  - ✓ Model selection, bayes estimators and risks

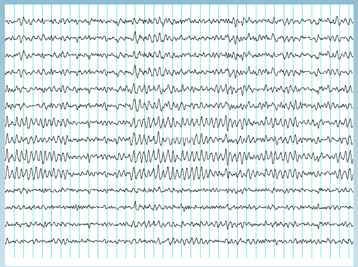
- Model adaptability
  - ✓ Markovian models, multi-streams HMM, structure adaptation, Markov random fields and CRF
  - ✓ Learning with unknown or evolutive costs, multi-objective learning, hyper-parameters in classifier ensembles (random forests, DRF, one-class)
  - ✓ Multi-task learning

Application domains

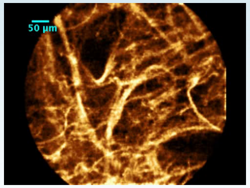
- Access to information
  - ✓ Handwriting recognition
  - ✓ Spotting
  - ✓ Information extraction
  - ✓ Complex manuscripts
  - ✓ Digital libraries
  - ✓ Recommandation systems

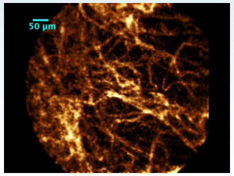


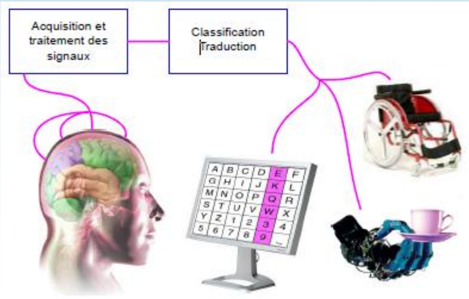




- Biomedical information processing
  - ✓ Brain Computer interface
  - ✓ Analysis of motor control data
  - ✓ Medical image classification
  - ✓ Medical image segmentation







STIC-AmSud French Team

- Prof. L. Heutte, PhD, PhD supervisor
  - ✓ Off-line and on-line handwriting analysis and recognition
  - ✓ Handwritten document analysis (bank checks, postal addresses, incoming mails, old manuscripts)
  - ✓ Information extraction and retrieval in handwritten documents
  - ✓ Classifier ensemble learning, classifier selection in ensembles
  - ✓ Pattern spotting,document image retrieval

- Ass. Prof. Caroline Petitjean, PhD
  - ✓ Medical image analysis, segmentation and classification
  - ✓ Cardiac MRI image segmentation with shape prior (graph-cut)
  - ✓ Medical image modelling
  - ✓ Image retrieval

- Ass. Prof. Simon Bernard, PhD
  - ✓ Classifier ensemble learning
  - ✓ Random forests

Random Forests

- A joint work with Simon BERNARD (PhD)
  - ✓ Understand why and how performance is affected by hyper-parameters
  - ✓ Improve Random Forest induction procedure
  - ✓ Forest-RK, Dynamic Random Forests

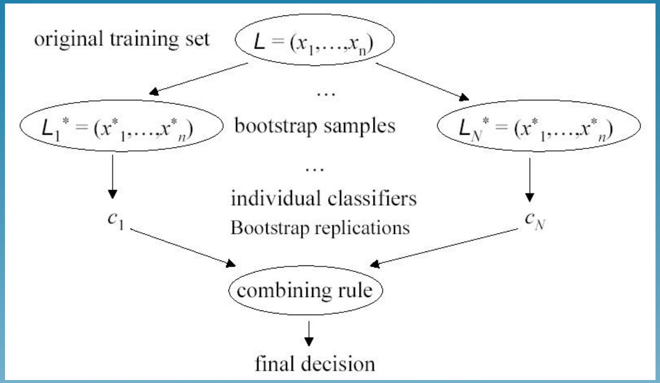
- A joint work with Chesner DESIR (PhD), Caroline PETITJEAN and Simon BERNARD
  - ✓ A new challenging application: endomicroscopic image classification
  - ✓ Optimizing Extra-Trees and random subwindows for medical image classification
  - ✓ One-Class Random Forest

Ensemble of Classifiers

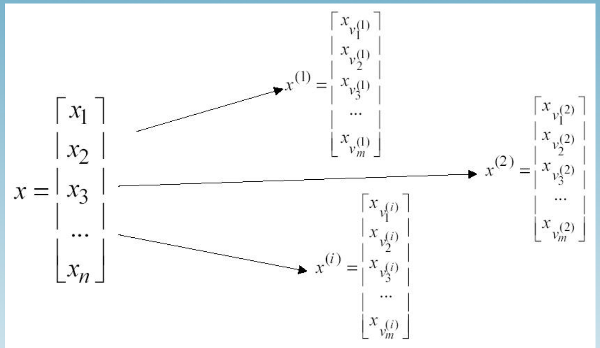
- IDEA:
  - ✓ Do not learn a *single classifier* but learn a *set of classifiers*
  - ✓ *Combine the predictions* of multiple classifiers
- MOTIVATION:
  - ✓ <u>Reduce variance</u>: results are less dependent on peculiarities of a single training set
  - ✓ <u>Reduce bias</u>: a combination of multiple classifiers may learn a more expressive concept class than a single classifier
- KEY STEP:
  - ✓ Building an ensemble of *diverse* classifiers from a single training set

Building an Ensemble of Classifiers

- By manipulating training data
  - ✓ cross-validated committees
  - ✓ bagging
  - ✓ Boosting



- By manipulating input features
  - ✓ adding noise
  - ✓ random subspace method



- By manipulating output features
  - ✓ error-correcting output codes

Exhaustive ECOC for $c = 4$ classes ($L = 7$ classifiers)

	D_1	D_2	D_3	D_4	D_5	D_6	D_7
ω_1	0	0	0	1	0	1	1
ω_2	0	0	1	0	0	0	0
ω_3	0	1	0	0	1	0	1
ω_4	1	0	0	0	1	1	0

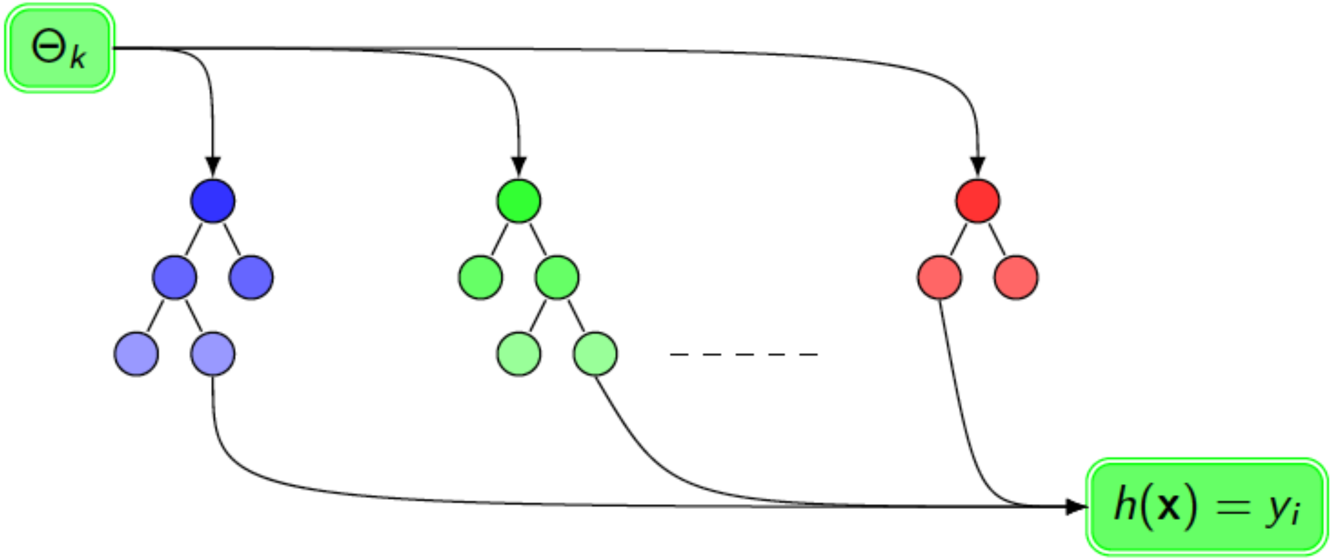
Random Forests [Breiman, 2001]

Definition

A random forest is a classifier consisting of a collection of tree-structured classifiers

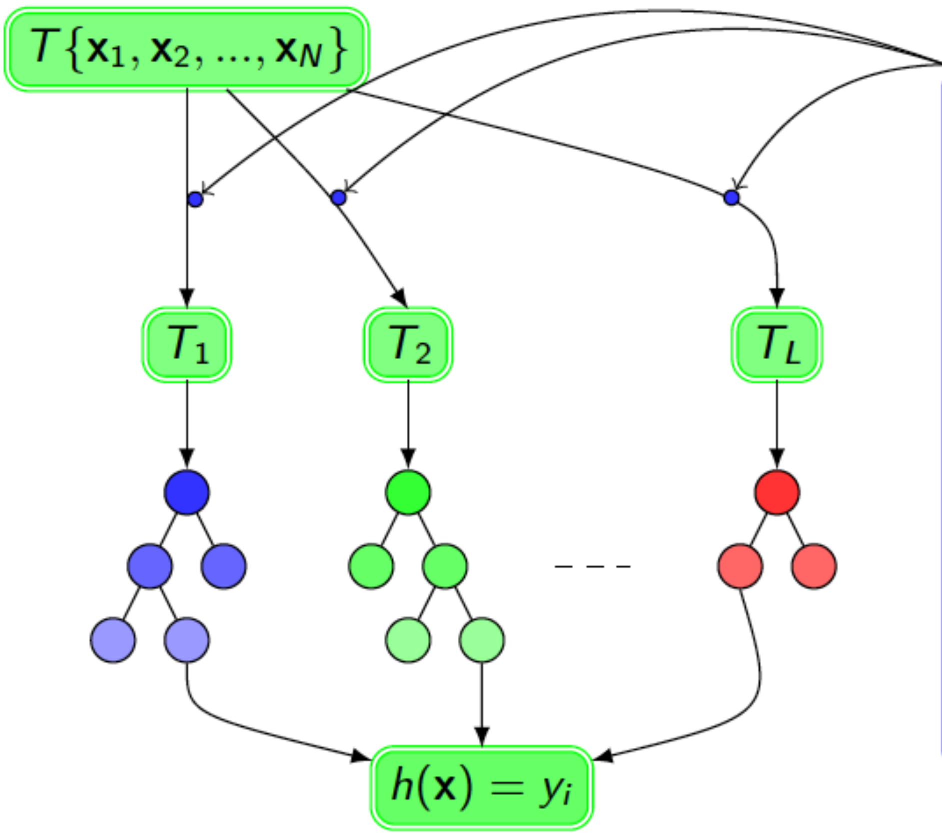
$$\{\ h(\mathbf{x}, \Theta_k), \quad k = 1, ..., L\ \}$$

where the $\{\Theta_k\}$ are independent identically distributed random vectors.



Forest-RI [Breiman, 2001]

Two randomization principles :

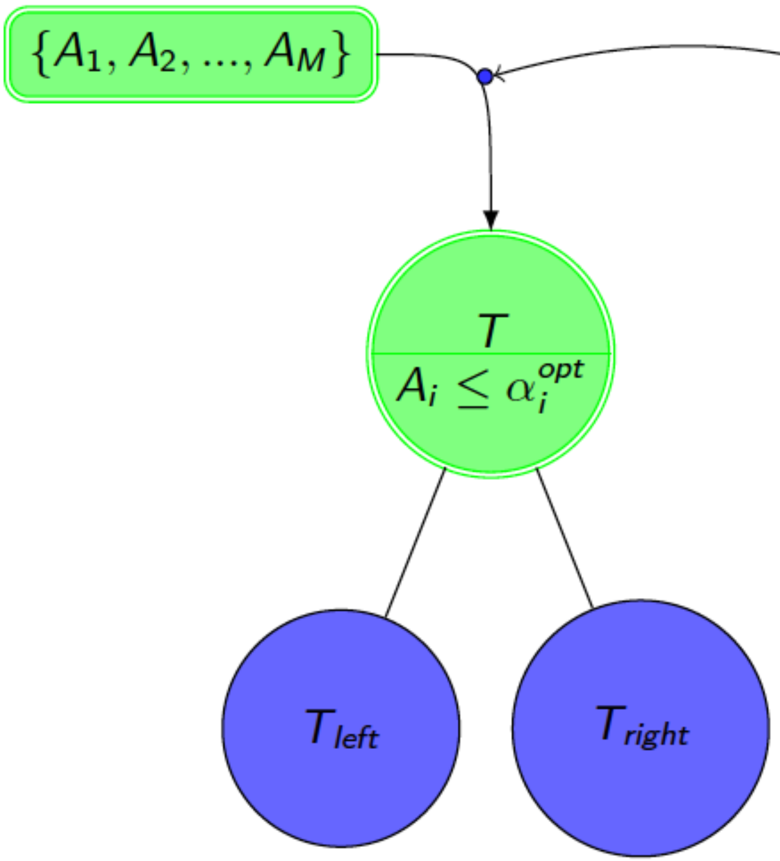


Bagging

- Random sampling with replacement of N training instances to form the boostra samples T_k, for $k = 1..L$
- the h_k classifier is trained on T_k
- aggregation of the L resulting classifiers for the prediction

Forest-RI [Breiman, 2001]

Two randomization principles :



Random Feature Selection (RFS)

- randomly choose K features among M
- built the best splitting rule from these features

Random Forests

- Random Forests are efficient...
  - ✓ Variable selection/variable importance (efficient, even in high dimensional feature space)
    - [Breiman, 2001], NIPS 2003 Workshop on Feature Extraction
  - ✓ "outlier" detection, proximity measures
    - [Breiman, 2001, Robnik-Sikonja, 2004]
  - ✓ Forest-RI comparable to AdaBoost
    - [Breiman, 2001, Baneld et al., 2004, Boinee et al., 2005, Geurts et al., 2006, Baneld et al., 2006]

- But this efficiency is not always easy to reach in practice...

An example…

- Madelon dataset (UCI repository, 2600 samples, 500 features, 2 classes)
  - ✓ Forest-RI : K = 22 (sqrt(M)), L = 300 => test error rate = 30.50%
  - ✓ Forest-RI : K = 260, L = 300 => test error rate = 17.73%
  - ✓ Forest-RI : K = 260, L = 100 (selected trees) => test error rate = 15.96%

- Understand why and how performance is affected by the hyper-parameters K and L
- Improve Random Forest induction procedure
- => PhD work of Simon Bernard

Influence of K: strength of randomization

- K controls the strength of randomization in the tree induction
  - ✓ Integer between 1 and M
  - ✓ K=1 => maximun randomization
  - ✓ K=M => no randomization (like CART, C4.5…)

- How to choose K?
  - ✓ Arbitrary values of K
  - ✓ K = 1, [Beiman 2001]
  - ✓ $K = \log_2(M)+1$, [Breiman 2001]
  - ✓ K = sqrt(M), [Geurts et al, 2006]

- Experiments on 14 UCI datasets with exhaustive search of K*
  - ✓ K* vs K_1, K_{sqrt}, K_{log} => K_{sqrt} is the closest to K* on average…
  - ✓ … but sometimes K_{sqrt} is a really bad choice!

Forest-RK: a new push-button algorithm

- Same as Forest-RI, i.e. with bagging and RFS…
- … but K is chosen randomly at each node
- => K is not an hyper-parameter anymore

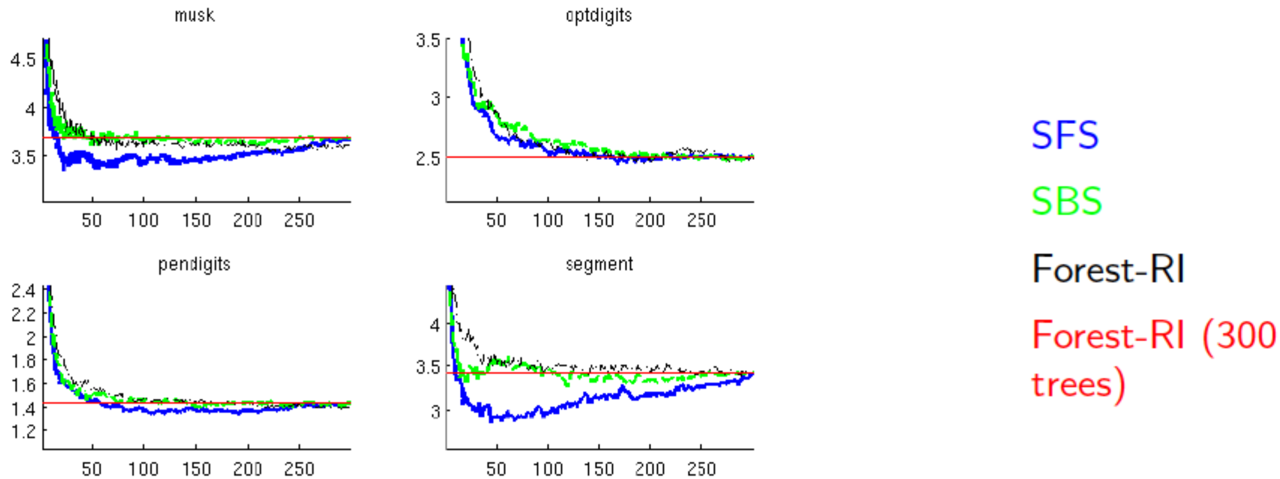
- Experiments conducted on 14 UCI datasets [MCS2009]
  - ✓ Forest-RK close to K* and at least as good as Ksqrt in majority

- Why does it work?
  - ✓ K acts as a trade-off for balancing performance and diversity
  - ✓ few relevant features => RFS deteriorates the "performance/diversity" trade-off
  - ✓ lot of relevant features => weakens the RFS effects in split selection

Bernard et al. Influence of hyperparameters on random forest accuracy. MCS 2009

Influence of L: tree selection in RF

Idea : Generate sub-forests and study their performance



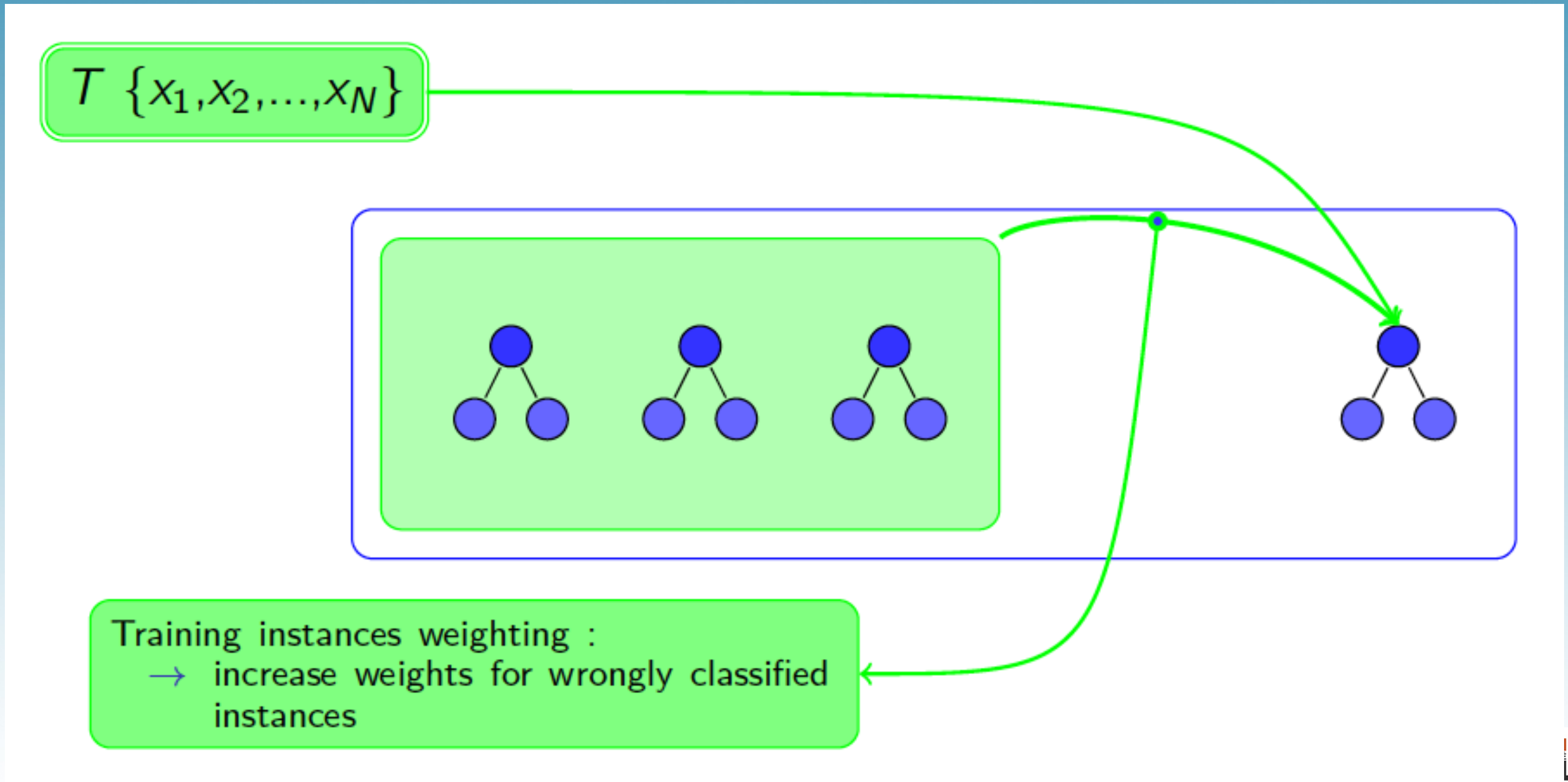
error rates with respect to the number of trees in sub-forests

→ **For** 18 **of the** 20 **datasets : at least one sub-forest significatively better**
→ **Sub-forests at least comparable to the 300-trees RF, with sometimes less than** 10% **of trees**

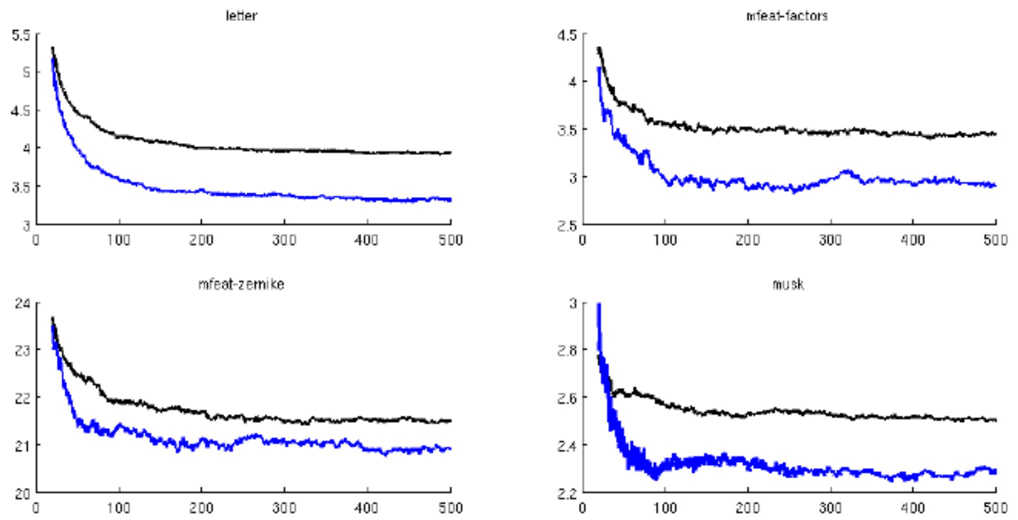
Bernard et al. On the selection of decision trees in random forests. IJCNN, 2009.

Dynamic Random Forests

- Main idea: guide the tree induction
  - ✓ make the new tree dependent on the subforest already built
  - ✓ Boosting-like procedure



Dynamic Random Forests



DRF

Forest-RI

Test error rates with respect to the number of trees

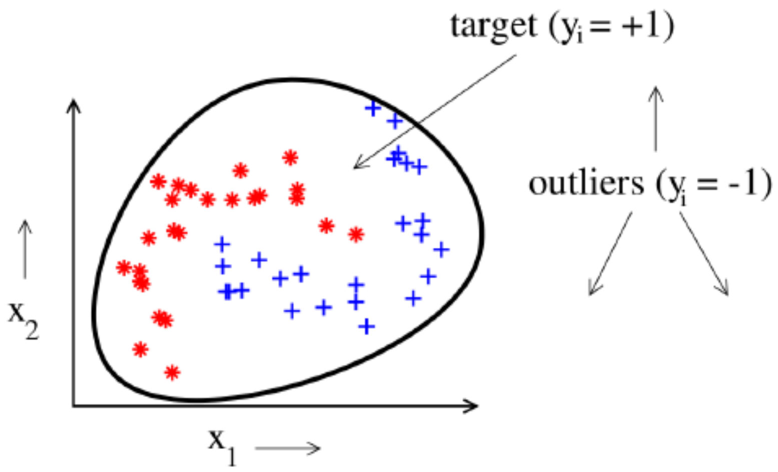
1. 500-trees forests
   → DRF significatively outperforms Forest-RI$/K_{sqrt}$ for 14 datasets
   → DRF $>$ *Forest-RI$/K^*$* $>$ Forest-RI$/K_{sqrt}$

- Bernard et al. Dynamic Random Forests. PRL, 33(12), 2012

One-Class Classification

- One-class classification (**OCC**) consists in:

  1. learning from only one class of objects (the target class, e.g. medical abundant healthy cases) ...
  2. ... and discriminating this class of interest from one or several other classes of objects (the outlier class), with no prior knowledge about the outlier class (e.g. rare pathological cases)

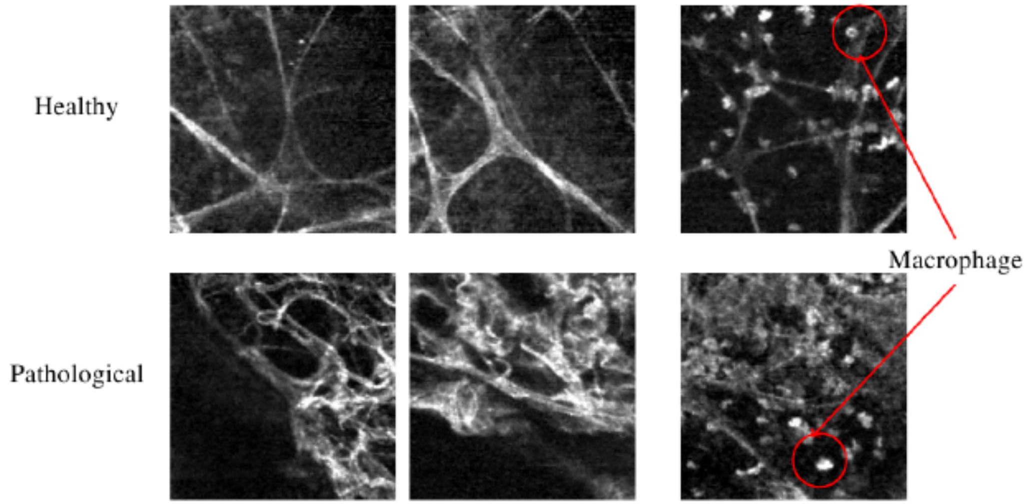


OCC is also called "novelty detection", "anomaly detection", "outlier detection" or "data description", often depending on applications.

One-Class Classification: Applications

Numerous applications:

- typist recognition, authorship verification, intrusion detection, mobile-masquerader detection, machine or structure health monitoring, fault detection...
- in the medical field: pathological samples are rare and therefore not reliable enough to train a binary classifier
- one example is the diagnosis of lung diseases via alveoscopy:



(a) Non-smoking subjects (b) Smoking subjects

Application in lung disease diagnosis

- *Classification of endomicroscopic images of the lung [MLMI 2012]*:
  - A new technique → new images → uncertain oracle on pathological images

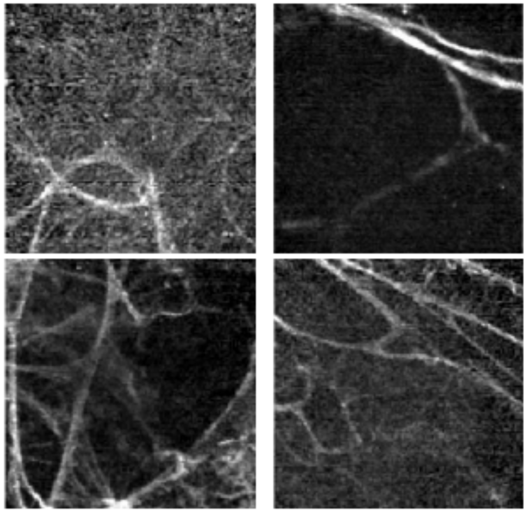


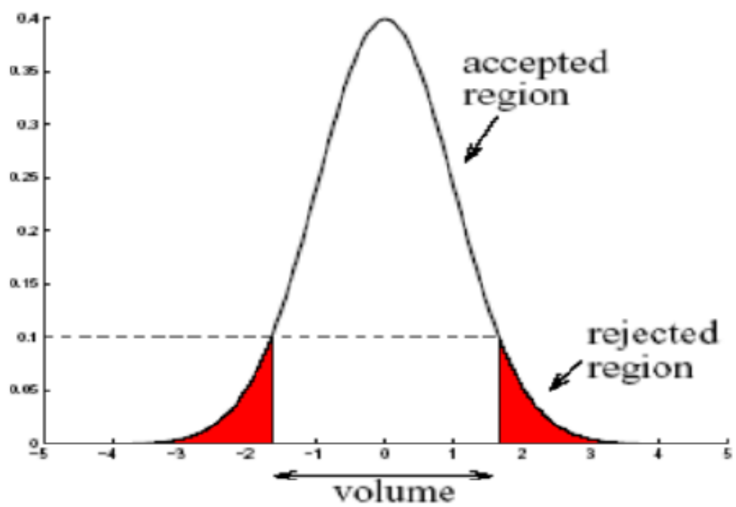
FIGURE 3 – Images alvéoscopiques présentant des difficultés de classification visuelle : sujet sain (haut), sujet pathologique (bas)

  - Learning from healthy images only → One-Class Paradigm

OCC: competitive algorithms (1)

- E

Natural choice for OCC [KM10]: Density estimators, based on estimating density function of target data: Mixture of gaussians, Parzen windows



Main issues with density approaches:

- Rarely effective for high dimensional problems, requiring large sample sizes for reliable estimation ("curse of dimensionality")
- Threshold on probability outputs chosen beforehand

OCC: competitive algorithms (2)

- Discriminative approaches, based on implicit decision frontier between classes to discriminate with
  - modification of inner workings of existing algorithms (SVDD , OCC-SVM)
    - SVDD [TD04] minimizes the volume of an hypersphere around the target class.
    - OCC-SVM [SPST$^+$01] separates the target from the origin with maximal margin with an hyperplane.
  - generation of artificial, uniformly distributed, outliers
    - hypercube of side *c* in *M*-sized feature space
    - requires an exponential amount of data to cover the whole feature space: **unusable**;

$$V_{hypercube}(c) = c^M$$

$$N_{outliers}(c) = \frac{V_{hypercube}}{V_{outlier}} = \frac{c^M}{(10^{-p})^M}$$

One-Class Random Forest

Our solution:

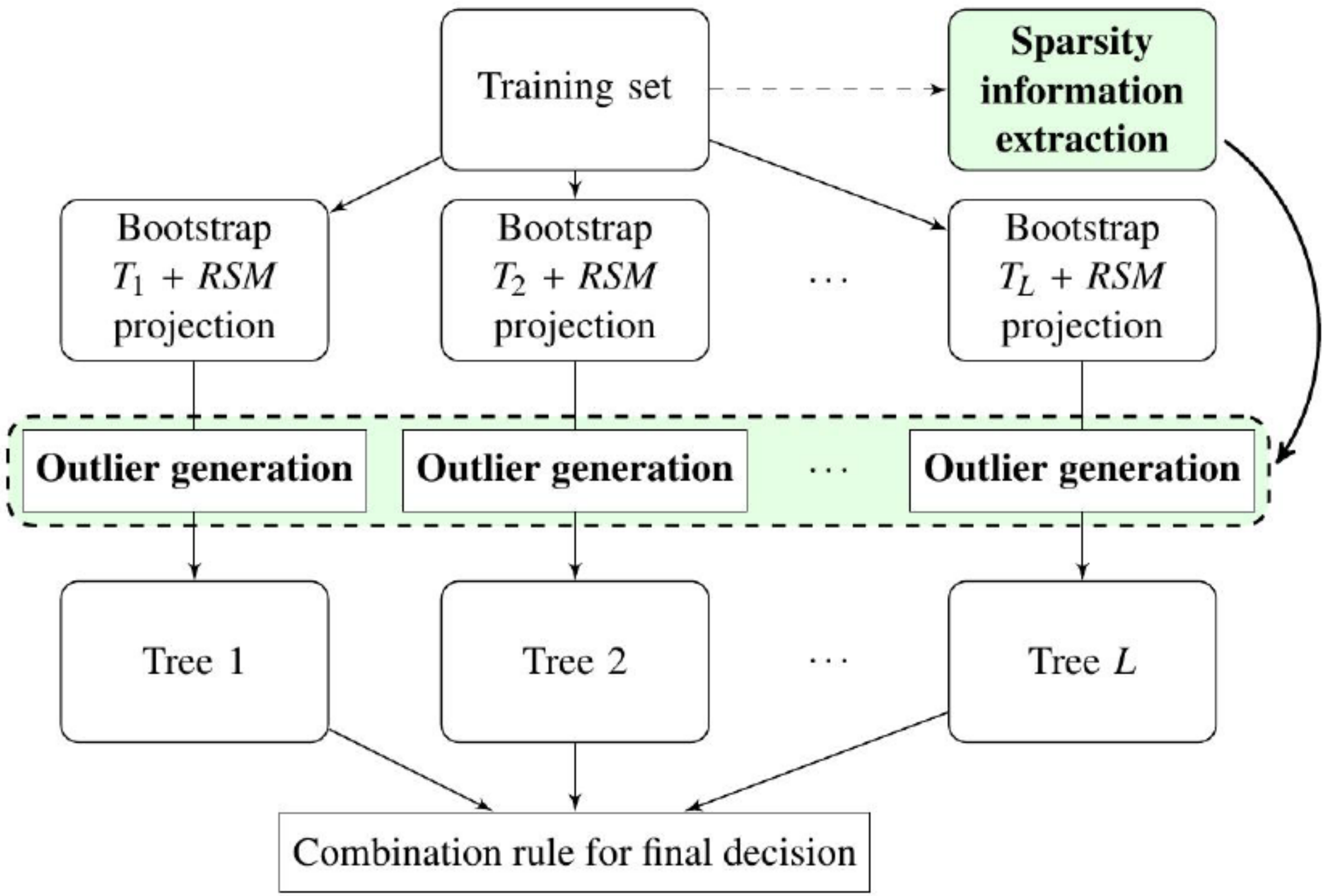
- Using classifier ensemble paradigm to break the curse of dimensionality, by subsampling the feature space and the training set.
- Generating more outlier data in sparse target regions, and less in densely populated target regions.

Our approach: One-Class Random Forest (**OCRF**), combining ensemble learning principles from traditional Random Forest algorithm with an original outlier generation method.

OCRF is composed of three main steps:

(i) extraction of **density information** from whole target data
(ii) generation of outlier data in **bootstrap** samples projected into **RSM** subspaces
(iii) induction of a **Forest-RI** on augmented dataset: bagging + RFS

OCRF: framework illustration



OCRF algorithm

Require: Training set T, $N_{outlier}$, $\Omega_{outlier}$, L individual trees, K_{RSM}

Ensure: A random forest classifier

(A) ***Density information extraction***
Compute H_{target}, normalized histogram of target data in T
Compute $H_{outlier}$, normalized histogram of generated outlier data, complementary of H_{target}, i.e. $H_{outlier} = 1 - H_{target}$

(B) ***Outlier generation and forest induction***
for $l = 1$ to L **do**
    (i) Draw a bootstrap sample T_l from training set T
    (ii) Project T_l onto a random subspace of dimension K_{RSM}
    (iii) Generate $N_{outlier}$ outlier data according to $H_{outlier}$ in the domain $\Omega_{outlier}$
    (iv) Train a standard decision tree on the augmented dataset
end for
return random forest model

OCRF algorithm

Advantages:

- full description obtained from whole dataset, instead of subsamples **(i)**
- far less artificial outlier data generated with RSM feature space sampling **(ii)** with a generation process adapted to the dataset **(i)**
- handling of high dimensional problems **(ii & iii)**
- general purpose framework, best or competitive results obtained with random-forest principles **(iii)**: bagging + RSM + RFS

Main parameters:

- Number of outliers to generate in **(ii)**
- Domain of outlier generation, e.g. surrounding hyperbox
- Forest parameters: L, K_{RSM}, K_{RFS}

OCRF experiments

- Results on 78 UCI datasets

Table: (a) Mean rank values; (b) Significancy results of statistical comparison with Friedman-Nemenyi test [Dem06]

	OCRF	OCSVM	Gauss	Parzen	Mog
Av. rank	2.4 ± 1.1	4.0 ± 1.3	1.9 ± 1.15	3.8 ± 1.1	2.8 ± 1.0

(a)

row $>$ col	OCRF	OCSVM	Gauss	Parzen	Mog
OCRF	-	+1	0	+1	0
OCSVM		-	-1	0	-1
Gauss			-	+1	+1
Parzen				-	-1

(b): OCRF $>$ {OCSVM,Parzen}, nothing can be said when compared to {Gauss} or {MoG}

Désir et al. One-Class Random Forests. PR, 46(12), 2013

Publications

- On Random Forest
  - ✓ S. Bernard, S. Adam, L. Heutte. Dynamic random forests. Pattern Recognition Letters, vol. 33, no. 12, pp. 1580-1586, 2012.
  - ✓ S. Bernard, L. Heutte, S. Adam. A study of strength and correlation in random forests. 2010 International Conference on Intelligent Computing, ICIC 2010, Changsha, China, Communications in Computer and Information Science, Springer, vol. 93, pp. 186-191, 2010.
  - ✓ S. Bernard, L. Heutte, S. Adam. Towards a better understanding of random forests through the study of strength and correlation. 2009 International Conference on Intelligent Computing, ICIC 2009, Ulsan, Korea, LNAI 5755, Springer, pp. 536-545, 2009.
  - ✓ S. Bernard, L. Heutte, S. Adam. On the selection of decision trees in random forests. 2009 International Joint Conference on Neural Networks, IJCNN 2009, Atlanta, USA, IEEE Proceedings, pp. 302-307, 2009.
  - ✓ S. Bernard, L. Heutte, S. Adam. Influence of hyperparameters on random forest accuracy. 8th International Workshop on Multiple Classifier Systems, MCS 2009, Reykjavik, Iceland, LNCS 5519, Springer, pp. 171-180, 2009.
  - ✓ S. Bernard, L. Heutte, S. Adam. Forest-RK: A new random forest induction method. 2008 International Conference on Intelligent Computing, ICIC 2008, Shanghai, China, LNCS 5227, Springer, pp. 430-437, 2008.

Publications

- On One-Class Random Forests
  - ✓ C. Désir, S. Bernard, C. Petitjean, L. Heutte. One Class Random Forests. Pattern Recognition, vol. 46, no. 12, pp. 3490-3506, 2013.
  - ✓ C. Désir, S. Bernard, C. Petitjean, L. Heutte. A new random forest method for one class classification. IAPR International Workshop on Statistical Techniques in Pattern Recognition, SPR 2012, Hiroshima, Japan, G.L. Gimel' farb et al. (Eds.): SSPR & SPR 2012, LNCS 7626, pp. 282–290, 2012.
  - ✓ C. Désir, S. Bernard, C. Petitjean, L. Heutte. A random forest based approach for one class classification in medical imaging. Third MICCAI International Workshop on Machine Learning in Medical Imaging, MLMI 2012, Nice, France, F. Wang et al. (Eds.): MLMI 2012, LNCS 7588, Springer, Heidelberg, pp. 250-257, 2012.

- On application of Random Forests to Medical Image Classification
  - ✓ C. Désir, C. Petitjean, L. Heutte, M. Salaün, L. Thiberville. Classification of endomicroscopic images of the lung based on random subwindows and extra-trees. IEEE Transactions on Biomedical Engineering, vol. 59, no. 9, pp. 2677-2683, 2012.

Discussion…