A Novel Random Forest Dissimilarity Measure for Multi-View Learning

Hongliu Cao[1,2], **Simon Bernard**[1], Robert Sabourin[2], Laurent Heutte[1]

[1] LITIS, Université de Rouen Normandie, France
[2] LIVIA, École de Technologie Supérieure (ÉTS), Montreal, Canada

25th International Conference on Pattern Recognition
10-15 January 2021, Milan, Italy.

- Multi-view learning : an instance is described by Q different vectors.
- The task is to learn :

$$h : \mathcal{X}^{(1)} \times \mathcal{X}^{(2)} \times \cdots \times \mathcal{X}^{(Q)} \to \mathcal{Y}$$

- A mutli-view training set T is composed of Q training replicates noted :

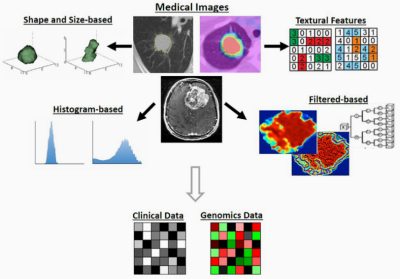
$$T^{(q)} = \left\{ (\mathbf{x}_1^{(q)}, y_1), (\mathbf{x}_2^{(q)}, y_2), \ldots, (\mathbf{x}_n^{(q)}, y_n) \right\}, \forall q = 1..Q$$

- Multi-view learning : an instance is described by Q different vectors.
- The task is to learn :

$$h : \mathcal{X}^{(1)} \times \mathcal{X}^{(2)} \times \cdots \times \mathcal{X}^{(Q)} \rightarrow \mathcal{Y}$$

- A mutli-view training set T is composed of Q training replicates noted :

$$T^{(q)} = \left\{ (\mathbf{x}_1^{(q)}, y_1), (\mathbf{x}_2^{(q)}, y_2), \ldots, (\mathbf{x}_n^{(q)}, y_n) \right\}, \forall q = 1..Q$$

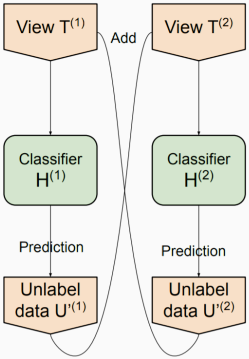


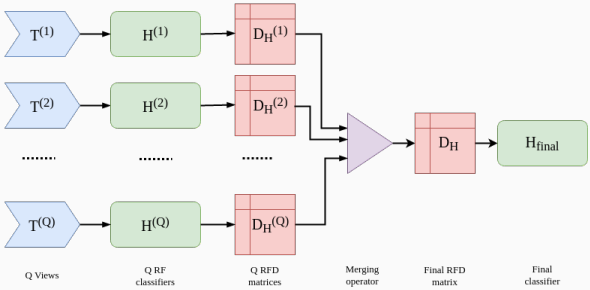
Example : Radiomics

- One or several modalities of medical images (CT, MRI, ...)
- Several families of features (Textures, Shapes/volumes,...)
- Combine with clinical and/or genomic data

Usually consists in learning separate models on each view and in ajusting them by maximing their agreement [10]

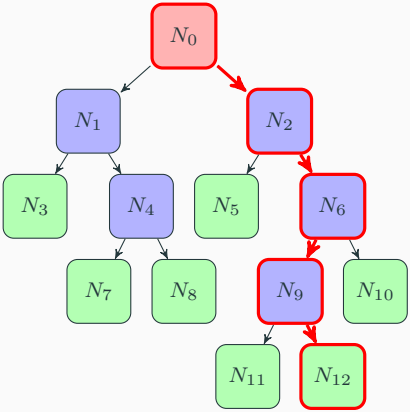
- For example, the most popular approach, Co-training methods [10]
- Problems : Require additional (unlabeled) data for adjusting the models
- This is often impossible for real-world problems for which data are particularly difficult to collect (e.g. medical field)





1. Train a Random Forest classifier $H^{(q)}$ on each $T^{(q)}$
2. From these RF, compute Q $n \times n$ dissimilarity matrices $D_H^{(q)}$, such that each cell is a dissimilarity measure $d(\mathbf{x}_i, \mathbf{x}_j)$ (more details after)
3. Merge the Q dissimilarity matrices to form a final RFD matrix D_H
4. Train a new classifier using D_H as a new training set

Random Forests embed a similarity measure on pairs of instances



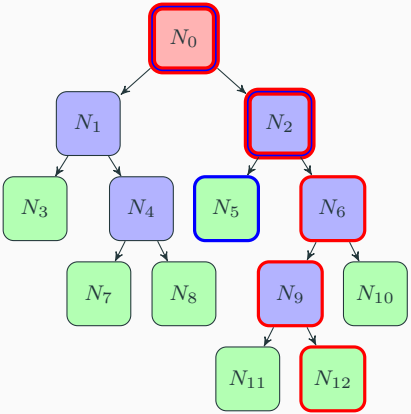
- Let $\mathcal{L}_k$ bet the set of leaves in the k^{th} tree of the forest
- Let

$$\ell_k : \mathcal{X} \to \mathcal{L}_k$$

be a function that maps all $\mathbf{x}$ to predict with that tree to the leaf from $\mathcal{L}_k$ in which it lands

- Here, $\ell_k(\mathbf{x}_i) = N_{12}$

Random Forests embed a similarity measure on pairs of instances



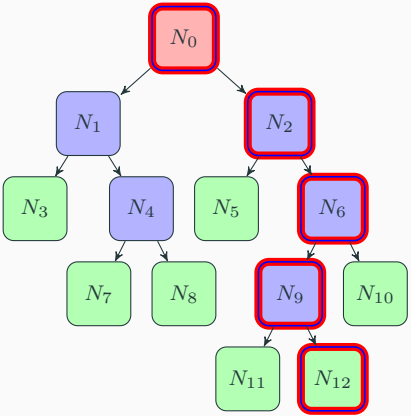
- The similarity $d^{(k)}(\mathbf{x}_i, \mathbf{x}_j)$ between $\mathbf{x}_i$ and $\mathbf{x}_j$, given by the k^{th} tree, is

$$d^{(k)}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1 & if\ \ell_k(\mathbf{x}_i) = \ell_k(\mathbf{x}_j) \\ 0 & otherwise \end{cases}$$

- Here, $\mathbf{x}_i$ and $\mathbf{x}_j$ don't land in the same leaf :

$$d^{(k)}(\mathbf{x}_i, \mathbf{x}_j) = 0$$

Random Forests embed a similarity measure on pairs of instances



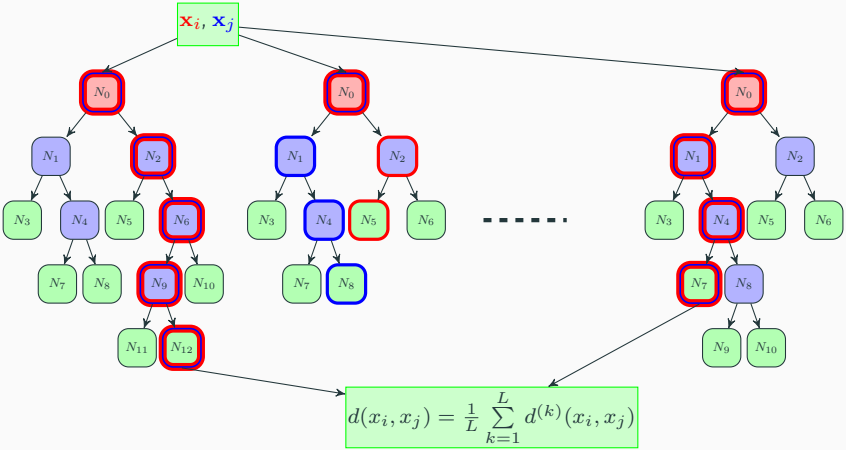
- The similarity $d^{(k)}(\mathbf{x}_i, \mathbf{x}_j)$ between $\mathbf{x}_i$ and $\mathbf{x}_j$, given by the k^{th} tree, is

$$d^{(k)}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1 & if\ \ell_k(\mathbf{x}_i) = \ell_k(\mathbf{x}_j) \\ 0 & otherwise \end{cases}$$

- Here, $\mathbf{x}_i$ and $\mathbf{x}_j$ land in the same leaf :

$$d^{(k)}(\mathbf{x}_i, \mathbf{x}_j) = 1$$

Random Forests embed a similarity measure on pairs of instances



The Random Forest Similarity measure pros and cons

Pros :

- Good theoretical properties ([3, 8])
- Non-parametric
- Take the class into account for learning the similarities
- No formulation of the metric beforhand (contrary to metric learning methods)

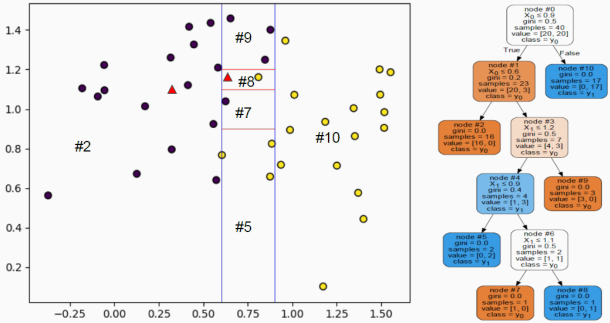
Cons :

- The tree-based measure is overly simplistic (0/1), which could lead to inaccurate measurement if the forest is composed of too few trees ([6])

⇒ We propose 2 new methods for measuring similarities with RF within the RFD framework

Proposed method 1 : RFD with Node Confidence (RFD_{NC})

All the leaves of a tree are not equally reliable for estimating similarities



According to the RF similarity measure :

- the 'red' instance in node #2 is similar to all the 'blue' instances in the same area
- the 'red' instance in node #8 is similar only to the 'yellow' instance in the same node

Proposed method 1 : RFD with Node Confidence (RFD_{NC})

All the leaves of a tree are not equally reliable for estimating similarities

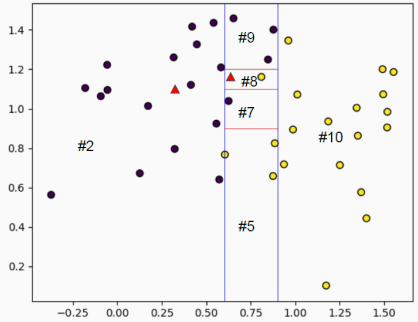
- Solution : Weight the RFD measure with a node confidence estimate
- Use Out-of-Bag instances ([1]) of each tree for computing these weights
- For a given instance $\mathbf{x}_t$, its weight is given by :

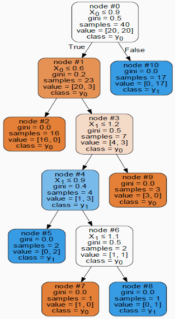
$$w_p(\mathbf{x}_t) = \frac{1}{|l_p(\mathbf{x}_t)|} \sum_{\mathbf{x}_i \in l_p(\mathbf{x}_t)} I(h_p(\mathbf{x}_i) = y_i)$$

where $|l_p(\mathbf{x}_t)|$ is the number of training instances, including the OOB, that have landed in the same terminal node as $\mathbf{x}_t$.

Proposed method 1 : RFD with Node Confidence (RFD_{NC})

An instance shouldn't have the same similarity to all the training instances of the node in which it is located





According to the RF similarity measure :

- the 'red' instance in node #2 have the same similarity to all the 'blue' instances in the same node

Proposed method 2 : RFD with Instance Hardness (RFD_{IH})

An instance shouldn't have the same similarity to all the training instances of the node in which it is located

- Solution : Weight the RFD measures with an instance hardness estimate ([9])
- Use the k-Disagreeing Neighbors (kDN) measure :

$$kDN(\mathbf{x}_i) = \frac{|\mathbf{x}_j : \mathbf{x}_j \in kNN(\mathbf{x}_i) \cap y_j \neq y_i|}{k}$$

  where $kNN(\mathbf{x}_i)$ stands for the k nearest neighbors of $\mathbf{x}_i$

- The dissimilarity between any $\mathbf{x}$ and the training instance $\mathbf{x}_i$ is :

$$d_p(\mathbf{x}, \mathbf{x}_i) = \begin{cases} kDN(\mathbf{x}_i), & if\ l_p(\mathbf{x}) = l_p(\mathbf{x}_i) \\ 1, & otherwise \end{cases}$$

Experimental validation

- 15 real-world multi-view datasets (medical, image and text classification)
- 4 competitors for estimating dissimilarities within the RFD framework :
  - Euclidean distance
  - the LMNN metric learning method ([5])
  - the original RFD method (e.g. in [7])
  - the RFD variant proposed in [6] ($RFDis_{PB}$)
- 10 times stratified random split 50% training - 50% test
- 2 statistical tests of significance :
  - Nemenyi post-hoc test with Critical Differences (CD) ([4])
  - Pairwise analysis based on the Sign test, from the number of wins, ties and losses

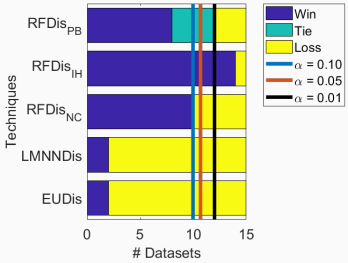
Average precision (with standard deviation) and mean rank

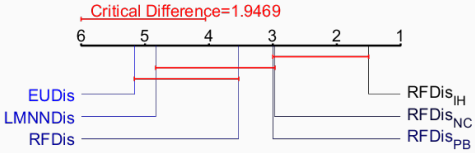
	$EUDis$	$LMNNDis$	$RFDis$	$RFDis_{PB}$	$RFDis_{NC}$	$RFDis_{IH}$
AWA8	39.22 ± 2.55	42.28 ± 3.13	56.06 ± 1.35	$\mathbf{56.38 \pm 1.47}$	56.34 ± 1.68	56.22 ± 1.01
AWA15	24.80 ± 0.97	28.25 ± 1.60	37.90 ± 1.49	37.62 ± 1.40	37.93 ± 1.50	$\mathbf{38.23 \pm 0.83}$
Metabo	$\mathbf{69.38 \pm 2.29}$	67.08 ± 4.04	67.71 ± 5.12	67.50 ± 5.76	67.08 ± 6.31	69.17 ± 5.80
Mfeat	96.00 ± 1.45	96.87 ± 0.79	97.56 ± 0.99	$\mathbf{97.63 \pm 0.95}$	97.63 ± 1.00	97.53 ± 1.00
NUS-WIDE2	89.52 ± 1.44	90.33 ± 1.55	92.49 ± 2.01	92.49 ± 1.81	92.67 ± 1.47	$\mathbf{92.82 \pm 1.93}$
BBC	85.89 ± 1.33	93.02 ± 1.29	92.82 ± 0.67	93.00 ± 0.67	92.33 ± 0.49	$\mathbf{95.46 \pm 0.65}$
lowGrade	63.72 ± 5.12	62.33 ± 7.04	63.48 ± 3.76	63.72 ± 4.67	$\mathbf{63.95 \pm 3.64}$	63.95 ± 5.62
NUS-WIDE3	73.92 ± 2.40	78.02 ± 2.69	79.41 ± 1.94	79.64 ± 2.19	79.91 ± 2.14	$\mathbf{80.32 \pm 1.95}$
progression	58.42 ± 4.82	62.63 ± 5.86	63.42 ± 6.49	63.42 ± 7.48	63.95 ± 6.56	$\mathbf{65.79 \pm 4.71}$
LSVT	82.86 ± 2.11	$\mathbf{85.24 \pm 2.84}$	83.33 ± 3.97	82.70 ± 3.44	83.49 ± 3.56	84.29 ± 3.51
IDHCodel	73.53 ± 5.42	71.47 ± 2.30	76.47 ± 3.95	76.47 ± 4.16	76.18 ± 3.82	$\mathbf{76.76 \pm 3.59}$
nonIDH1	79.07 ± 3.45	73.26 ± 3.49	79.53 ± 3.57	79.53 ± 3.72	79.77 ± 3.46	$\mathbf{80.70 \pm 3.76}$
BBCSport	80.11 ± 1.69	73.77 ± 5.45	81.75 ± 2.70	82.56 ± 2.85	79.93 ± 3.11	$\mathbf{90.18 \pm 1.96}$
Cal20	84.04 ± 0.82	87.50 ± 0.78	89.12 ± 0.69	89.27 ± 1.01	89.06 ± 1.19	$\mathbf{89.76 \pm 0.80}$
Cal7	92.67 ± 0.63	95.09 ± 0.66	95.21 ± 0.67	95.51 ± 0.50	95.34 ± 0.48	$\mathbf{96.03 \pm 0.53}$
Avg rank	5.20	4.83	3.67	2.83	2.93	**1.53**

- $RFDis_{IH}$ is the most accurate method on 10 datasets. Its average rank is 1.53
- The RF-based dissimilarity methods achieve the best results for 14 datasets

Statistical significance

- These results are confirmed by the statistical tests





Conclusion

- RF measures are more accurate and better reflect the dissimilarities between instances with respect to the classification task, while remaining robust to high dimensions.
- The most efficient method is based on an instance hardness measurement calculated in the subspaces extracted from the trees of the RF.
- It allows to penalize unreliable dissimilarity estimates given by trees that have failed to correctly predict the instances.
- Experiments and results on real-world multi-view datasets have shown that this mechanism is significantly more accurate than the standard RFD measure and than state-of-the-art metric learning methods.

References

[1] Leo Breiman. “Random forests”. In : *Machine Learning* 45.1 (2001), p. 5-32.

[2] Hongliu Cao et al. “Random forest dissimilarity based multi-view learning for Radiomics application”. In : *Pattern Recognition* 88 (2019), p. 185-197.

[3] Alex Davies et Zoubin Ghahramani. “The random forest kernel and other kernels for big data from random partitions”. In : *arXiv preprint arXiv:1402.4293* (2014).

[4] Janez Demšar. “Statistical comparisons of classifiers over multiple data sets”. In : *Journal of Machine Learning Research* 7.Jan (2006), p. 1-30.

[5] Carlotta Domeniconi, Dimitrios Gunopulos et Jing Peng. “Large margin nearest neighbor classifiers”. In : *IEEE transactions on Neural Networks* 16.4 (2005), p. 899-909.

[6] Cristofer Englund et Antanas Verikas. “A novel approach to estimate proximity in a random forest : An exploratory study”. In : *Expert Systems with Applications* 39.17 (2012), p. 13046-13050.

[7] Katherine R Gray et al. “Random forest-based similarity measures for multi-modal classification of Alzheimer’s disease”. In : *NeuroImage* 65 (2013), p. 167-175.

[8] E. Scornet. “Random Forests and Kernel Methods”. In : *IEEE Transactions on Information Theory* 62.3 (2016), p. 1485-1500.

[9] Michael R Smith, Tony Martinez et Christophe Giraud-Carrier. “An instance level analysis of data complexity”. In : *Machine Learning* 95.2 (2014), p. 225-256.

[10] Jing Zhao et al. “Multi-view learning overview : Recent progress and new challenges”. In : *Information Fusion* 38 (2017), p. 43-54.