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, 76000 Rouen, France

[2]LIVIA, École de Technologie Supérieure (ÉTS), Université du Québec, Montreal, QC, Canada

Multi-view learning (MVL)

Instances are described by Q different vectors and the task is to learn:

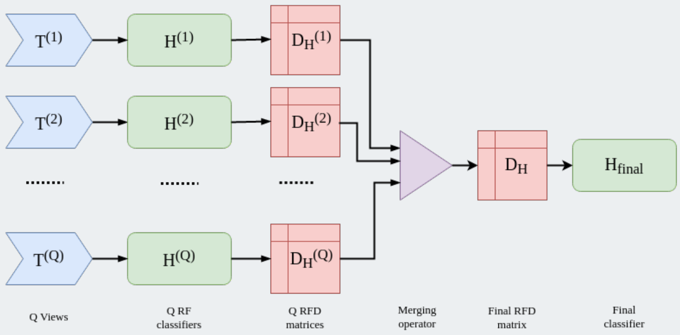
$$h : \mathcal{X}^{(1)} \times \mathcal{X}^{(2)} \times \cdots \times \mathcal{X}^{(Q)} \rightarrow \mathcal{Y}$$

A MVL training set T is typically composed of Q subsets:

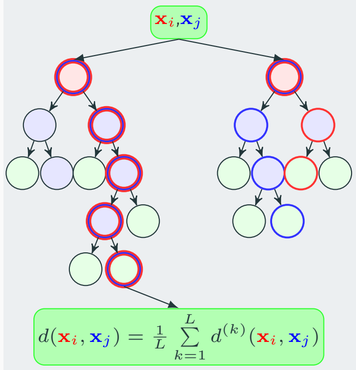
$$T^{(q)} = \left\{(\mathbf{x}_1^{(q)}, y_1), (\mathbf{x}_2^{(q)}, y_2), \dots, (\mathbf{x}_n^{(q)}, y_n)\right\}, \forall q = 1..Q$$

The Random Forest Dissimilarity (RFD) framework [2]

1. Compute Q $n \times n$ dissimilarity matrices from the $T^{(q)}, \forall q = 1..Q$ such that each cell is a dissimilarity $d(\mathbf{x}_i, \mathbf{x}_j)$
2. Dissimilarities are measured with a Random Forest (RF) trained on $T^{(q)}$
3. Merge the Q dissimilarity matrices to form a joint RFD matrix
4. Train a new classifier on this RFD matrix as a new training set



The RF dissimilarity measure



- Let $\mathcal{L}_k$ be the set of leaves in the k^{th} tree and
  $$l_k : \mathcal{X} \rightarrow \mathcal{L}_k$$
  be a function that maps any $\mathbf{x}$ to its leaf from $\mathcal{L}_k$
- The similarity $d^{(k)}(\mathbf{x}_i, \mathbf{x}_j)$ given by the k^{th} tree, is
  $$d^{(k)}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1, if\, l_k(\mathbf{x}_i) = l_k(\mathbf{x}_j) \\ 0, otherwise \end{cases}$$
- The similarity $d(\mathbf{x}_i, \mathbf{x}_j)$ given by the forest is the average of the $d^{(k)}(\mathbf{x}_i, \mathbf{x}_j)$ over all the trees
- The final dissimilarity is given by $1 - d(\mathbf{x}_i, \mathbf{x}_j)$

We argue that this measure is too rough (0/1), particularly for MVL

⇒ New method for measuring dissimilarity with RF for Multi-View Learning

Contributions

1. Use RF classifiers for learning dissimilarity representations for MVL
2. Two novel ways to learn dissimilarities from RF classifiers within the RFD framework
3. Validation by comparing them to 4 methods from the literature, including metric learning and other RF-based dissimilarity measure

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

- Issue: all the leaves are not equally reliable for estimating (dis)similarities
- Proposition:
  - 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 2 : RFD with Instance Hardness (RFD_{IH})

- Issue: an instance have the same dissimilarity to all the training instances of the node in which it is located
- Proposition:
  - Weight the RFD measures with an instance hardness estimate ([7])
  - 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 (see $EUDiss$ results in the paper)
  - the LMNN metric learning method ([4])
  - the original RFD method (e.g. in [6])
  - the RFD variant proposed in [5] ($RFDis_{PB}$)
- 10 times stratified random split 50% training - 50% test
- 2 statistical tests of significance:
  - Nemenyi post-hoc test with Critical Differences (CD) ([3])
  - Pairwise analysis based on the Sign test, from the number of wins, ties and losses

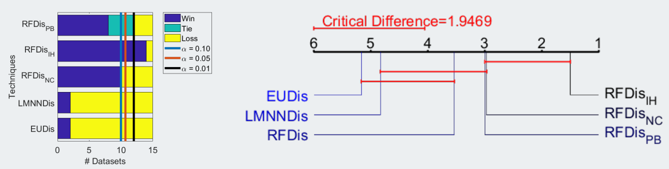
Results

Average precision (with standard deviation) and mean rank

	$LMNNDis$	$RFDis$	$RFDis_{PB}$	$RFDis_{NC}$	$RFDis_{IH}$
AWA8	42.28 ± 3.13	56.06 ± 1.35	$\mathbf{56.38 \pm 1.47}$	56.34 ± 1.68	56.22 ± 1.01
AWA15	28.25 ± 1.60	37.90 ± 1.49	37.62 ± 1.40	37.93 ± 1.50	$\mathbf{38.23 \pm 0.83}$
Metab.	67.08 ± 4.04	67.71 ± 5.12	67.50 ± 5.76	67.08 ± 6.31	69.17 ± 5.80
Mfeat.	96.87 ± 0.79	97.56 ± 0.99	$\mathbf{97.63 \pm 0.95}$	97.63 ± 1.00	97.53 ± 1.00
NUSW2	90.33 ± 1.55	92.49 ± 2.01	92.49 ± 1.81	92.67 ± 1.47	$\mathbf{92.82 \pm 1.93}$
BBC	93.02 ± 1.29	92.82 ± 0.67	93.00 ± 0.67	92.33 ± 0.49	$\mathbf{95.46 \pm 0.65}$
lowGr.	62.33 ± 7.04	63.48 ± 3.76	63.72 ± 4.67	$\mathbf{63.95 \pm 3.64}$	63.95 ± 5.62
NUSW3	78.02 ± 2.69	79.41 ± 1.94	79.64 ± 2.19	79.91 ± 2.14	$\mathbf{80.32 \pm 1.95}$
progr.	62.63 ± 5.86	63.42 ± 6.49	63.42 ± 7.48	63.95 ± 6.56	$\mathbf{65.79 \pm 4.71}$
LSVT	$\mathbf{85.24 \pm 2.84}$	83.33 ± 3.97	82.70 ± 3.44	83.49 ± 3.56	84.29 ± 3.51
IDHCo.	71.47 ± 2.30	76.47 ± 3.95	76.47 ± 4.16	76.18 ± 3.82	$\mathbf{76.76 \pm 3.59}$
nIDH1	73.26 ± 3.49	79.53 ± 3.57	79.53 ± 3.72	79.77 ± 3.46	$\mathbf{80.70 \pm 3.76}$
BBCSp.	73.77 ± 5.45	81.75 ± 2.70	82.56 ± 2.85	79.93 ± 3.11	$\mathbf{90.18 \pm 1.96}$
Cal20	87.50 ± 0.78	89.12 ± 0.69	89.27 ± 1.01	89.06 ± 1.19	$\mathbf{89.76 \pm 0.80}$
Cal7	95.09 ± 0.66	95.21 ± 0.67	95.51 ± 0.50	95.34 ± 0.48	$\mathbf{96.03 \pm 0.53}$
Avg rank	4.83	3.67	2.83	2.93	$\mathbf{1.53}$

Analysis

- $RFDis_{IH}$ 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
- These results are confirmed by the statistical tests (cf. Figure below)



Aknowledgment

This work is part of the DAISI project, co-financed by the European Union with the European Regional Development Fund (ERDF) and by the Normandy Region.

[1] Leo Breiman. "Random forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.
[2] Hongliu Cao et al. "Random forest dissimilarity based multi-view learning for Radiomics application". In: *Pattern Recognition* 88 (2019), pp. 185–197.
[3] Janez Demšar. "Statistical comparisons of classifiers over multiple data sets". In: *Journal of Machine Learning Research* 7 (2006), pp. 1–30.
[4] Carlotta Domeniconi, Dimitrios Gunopulos, and Jing Peng. "Large margin nearest neighbor classifiers". In: *IEEE transactions on Neural Networks* 16.4 (2005), pp. 899–909.
[5] Cristofer Englund and Antanas Verikas. "A novel approach to estimate proximity in a random forest: An exploratory study". In: *Expert Systems with Applications* 39.17 (2012), pp. 13046–13050.
[6] Katherine R Gray et al. "Random forest-based similarity measures for multi-modal classification of Alzheimer's disease". In: *NeuroImage* 65 (2013), pp. 167–175.
[7] Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier. "An instance level analysis of data complexity". In: *Machine Learning* 95.2 (2014), pp. 225–256.