NEW ITERATIVE LEARNING STRATEGY TO IMPROVE CLASSIFICATION SYSTEMS BY USING OUTLIER DETECTION TECHNIQUES

C. Pelletier, S. Valero, J. Inglada, G. Dedieu

CESBIO - UMR 5126
18 avenue Edouard Belin
31401 Toulouse CEDEX 9 - FRANCE

N. Champion

IGN Espace - MATIS / Université Paris-Est
6 avenue de l'Europe
31521 Ramonville CEDEX - FRANCE

ABSTRACT

The supervised classification of satellite image time series allows obtaining reliable land cover maps over large areas. However, their quality depends on the reference datasets used for training the classifier. In remote sensing, reference data may lack of timeliness and accuracy which leads to the presence of mislabeled data degrading the classification performances. This work presents an iterative learning framework to deal with noisy instances, that can be seen as outliers. Several outlier detection strategies, based on the well-known Random Forests (RF) ensemble classifier, are proposed, evaluated quantitatively, and then compared with traditional methods. Experimental results have been carried out by using synthetic and real datasets representing annual vegetation profiles.

Index Terms— mislabeled data, outlier detection, Satellite Image Time Series classification, Random Forests, land cover mapping

1. INTRODUCTION

The supervised classification of new satellite image time series (SITS) such as Sentinel-2 allows the production of accurate land cover maps over large areas [1]. In practice, supervised learning algorithms require a subsequent number of well-labeled data to train the classifier model and to evaluate the land cover product.

In remote sensing, such reference datasets can come from different sources such as field measurements, thematic maps, or aerial photographs. Accordingly, they may contain class label noise, *i.e.* instances with a wrong label assignment, due to misregistration, land cover complexity, or update delay. For instance, the 2012 version of Corine Land Cover (CLC) produced at European scale was made available only in 2015. If this data is used to train a classifier aiming at classifying a SITS captured at 2015, the training set will fatally contain mislabeled data leading to some training errors.

The presence of mislabeled data degrades the classification performances [2]. A solution to handle the mislabeled data problem is the use of active learning strategies, where users select iteratively the best informative instances [3, 4]. Unfortunately, these methods are hardly applicable when a large reference dataset is needed. Therefore, outlier detection methods can be a solution to improve the classification performances by filtering [5, 6] or by correcting the mislabeled data [7]. In this work, a new strategy to tackle the mislabeled data detection problem is proposed. The goal is to remove iteratively the outlier instances from the training dataset.

Traditional outlier detection methods search for rare instances in a dataset that are significantly different from the remainder of the instances. Most of the literature methods are distance-based approaches such as k-Nearest Neighbor (kNN) [8], isolation forest (iForest) [9] and Local Outlier Factor (LOF) [10]. Distance-based approaches consider an instance as an outlier if it does not exist similar instances in a pre-defined neighborhood. Therefore, their main drawback is the difficulty to determine an appropriate value for the user-defined parameter k that sets the neighborhood size. In addition, these methods are generally based on the Euclidean distance, which is not an appropriate similarity measure for instances described by time series [11].

Besides, the analysis of SITS implies to handle high dimensional data, where traditional outlier detection approaches can fail. In this context, the well-known Random Forests (RF) classifier has shown its interest [12]. The RF tree structure allows the computation of similarities between pairs of instances in high dimensional datasets. More precisely, the similarity measure proposed by Breiman counts for the number of times that instances fall into the same terminal node. An outlier score computed from the similarities can be then used to detect outlier instances. In this work, this outlier detection strategy is studied in an iterative classification framework where the instances with the highest outlier scores are removed. In practice, this similarity measure can be coarse. Thus, new measures also based on the RF tree structure are proposed in this study.

Hence, this work aims at presenting a new iterative learning strategy by studying reliable outlier detection methods. To assess the proposed strategy, studies are carried out on synthetic and real datasets describing vegetation profiles over one year. First, the effectiveness of outlier detection methods by using the RF tree structure is evaluated quantitatively, and compared with traditional methods such as kNN, LOF and iForest. A second experiment corroborates the interest of the proposed iterative learning procedure for improving the classification performances in the presence of mislabeled training data.

The remainder of this paper is organized as follows. The proposed methods are described in details in Section 2. Then, Section 3 introduces the data used in this work. Section 4 provides some experimental results yielded by the proposed approaches and compared with some reference methods. Eventually, conclusions are drawn in Section 5.

2. METHOD

2.1. Iterative learning procedure

In this work, a new learning strategy that aims at removing iteratively mislabeled training instances is proposed. Fig. 1 describes the proposed framework. It consists in learning a RF model, computing an outlier score, and removing from the training set the top n rank out-

This work is funded by the French spatial agency (CNES) and the French mapping agency (IGN).

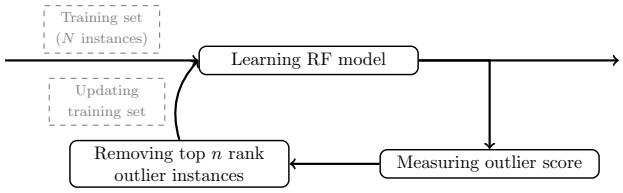

Fig. 1. Proposed iterative learning procedure.

liers, *i.e.* the n instances with the highest outlier scores. These three steps are iteratively repeated. In real application, the noise level differs for each class. Thus, the outlier measures are ranked regardless of their class, *i.e.* the n removed instances may belong to the same class.

2.2. Outlier detection strategy using Random Forest trees

The RF classifier is an ensemble learning method that consists in learning a set of weak classifiers (decision trees) to generate a classifier with a strong decision rule. Each tree is formed by selecting at random, at each node, a number of features to split on.

Breiman proposes the computation of an outlier score for each instance by using the set of the trees built by the RF classifier. The higher outlier score O is, the higher probability that the instance is mislabeled. To compute it, the similarity between each pair of instances from the class is evaluated. After each tree is built, all of the data run down the tree until the terminal nodes. Then, the similarity between the ith and jth instances is computed as the fraction of trees in which both instances fall in the same terminal node. The main assumption is that similar instances should be in the same terminal nodes more often than dissimilar ones.

Accordingly, the raw outlier measure O_{raw} for the ith instance is defined as:

$$O_{raw}(i) = \frac{n_c - 1}{\sum\limits_{j=1, j\neq i}^{n_c} P(i,j)^2}, \quad (1)$$

where n_c denotes the number of samples belonging to the class of the ith instance, and $P(i,j)$ the similarity (also called proximity) between the ith and jth instances. To compare the outlier scores between classes, med_c the median of O_{raw} measures and MAD_c the median absolute deviation are used to scale the raw outlier measures per class:

$$O(i) = \frac{O_{raw}(i) - med_c}{MAD_c}. \quad (2)$$

The similarity measure proposed by Breiman can be seen as a rough binary measure. For one tree, the proximity between two instances equals zero or one. Accordingly, this measure requires many trees to get a stable estimation of the proximities [13].

2.3. New proximity measures computed by using Random Forest trees

Two new proximity measures are presented here to improve the classical $Binary$ measure. More precisely, they consider the number of edges that separates the terminal nodes where the instances fall. $NbEdges$ counts the number of edges in the trees that separates the instances, and $MaxAncestor$ counts the maximum number of edges that separates the given instances from their common ancestor.

3. STUDIED DATASET

3.1. Datasets

The experiments are carried out on synthetic and real datasets representing Normalized Difference Vegetation Index (NDVI) profiles. Both datasets cover one year and describe five vegetation classes.

The synthetic NDVI profiles are generated by the double logistic equation described in [1]. For each class, six parameters are used to describe the NDVI profiles. Each simulated profile is composed of 15 dates. To create realistic profiles, a vegetation regrowth and a uniform noise have also been added to the profiles. A complete description of the generation procedure will be soon available in the submitted paper [14].

The real dataset is obtained by extracting NDVI profiles from 23 images captured on 2013 by SPOT-4 and Landsat-8 satellites in the Southwest of France. The reference data are extracted from the French Land Parcel Information System, that annually maps a set of polygons describing the French crop fields.

Both datasets are divided in two sets: one for training, and one for validation. Each set is composed of 500 instances per class.

3.2. Label noise generation procedure

To evaluate quantitatively the outlier detection methods, artificial mislabeled data is injected in the training datasets. This step consists in modifying randomly the label of some instances. The wrong label assignment is equiprobable between all the class labels except the original one. The same noise level is injected for each class. A total of nineteen noise levels, ranging from 5 % to 95 % with 5 % step, are studied.

4. EXPERIMENTAL RESULTS

4.1. Quality assessment of outlier detection methods

The outlier detection methods presented in Section 2 are evaluated quantitatively and compared with the following traditional methods: kNN, LOF, iForest, Edited Nearest Neighbor (ENN) [5] and All$k$NN [15]. In this study, the specificity of remote sensing data – reference datasets are usually composed of polygons – has been taken into account. The proposed processing consists in computing similarity measures only between instances that do not belong to the same polygon. The goal is to avoid the use of correlated instances, *i.e.* belonging to the same polygon, to compute the outlier scores. This processing has been implemented for the literature and the proposed methods (except iForest).

The quality assessment of the methods is evaluated by using three common criteria: the Area Under the Curve of the Receiver Operating Characteristic (ROC-AUC) curve, the Average Precision (AP) [16], and the F-Score values [17]. Concerning the ENN and All$k$NN methods, only the F-Score evaluation is performed since both methods do not compute an outlier score for each instance.

The synthetic and real datasets composed of 500 instances per class (Section 3.1) are used. The presented noise injection procedure is applied for nineteen noise levels ranging from 5 % to 95 % with 5 % step.

Fig. 2 displays ROC-AUC, AP and F-Score values as a function of the noise levels. The first row displays the results for synthetic data, and the second row for real data. Each curve color displays the results obtained by a specific method. The best k parameter values is selected at each noise level for kNN, LOF, ENN and All$k$NN methods. The number of trees used in iForest and RF algorithms is equal to 100.

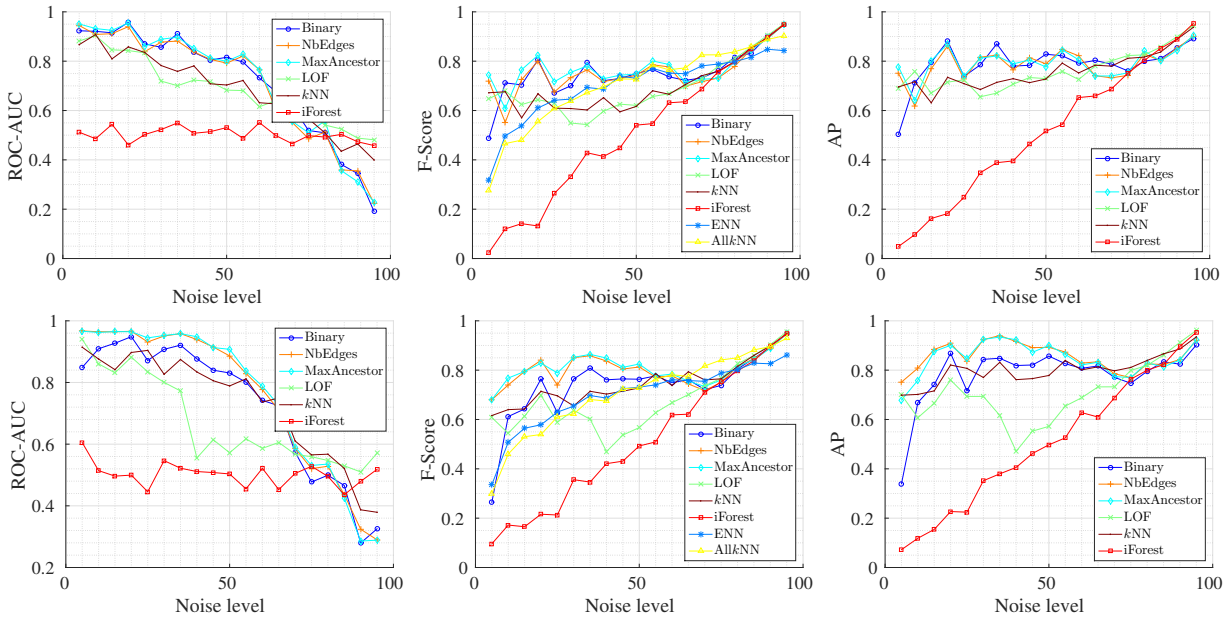

Fig. 2. ROC-AUC, F-Score, and AP as a function of noise level for different outlier detection methods. First row: 5-class synthetic dataset. Second row: 5-class real dataset.

These results show that the outlier detection methods using RF trees obtain better detection performances than traditional methods for all the evaluation criteria. They also show that Euclidean distance computed by classical methods fails with high dimensional remote sensing data. It may be noticed that the proposed $NbEdges$ and $MaxAncestor$ proximity measures overcome the classical $Binary$ proximity measure proposed by Breiman.

4.2. Evaluation of the proposed iterative learning procedure

The iterative learning strategy presented in Section 2.1 is evaluated by using the outlier detection methods whose similarity measures are based on the RF trees. This iterative classification procedure requires the setting of the parameter n. Preliminary results are displayed here. Consequently, a small value for n (equals to 10) is chosen to assure the removal of only outliers. In this experiment, the behavior of the proposed system is analyzed for the first iterations. A stopping criteria will be defined in future works.

The evaluation is carried out by using the Overall Accuracy (OA) computed at each iteration. The validation set used is composed of 500 training instances per class and free of noise. A second evaluation is performed by studying the cumulative number of removed outliers at each iteration, i.e. the number of mislabeled data among the total number of removed instances.

Fig. 3 shows the results obtained on the synthetic dataset. The first row displays the results obtained for 20 % noise level, and the second row for 40 % noise level. The first column shows the evolution of OA through the different iterations, and the second column shows the cumulative number of removed outliers at each iteration. Two important stages are highlighted on the OA figures: (1) the horizontal red dashed line which represents the OA values obtained if all the mislabeled data were corrected, and (2) the horizontal purple dashed line which represents the OA accuracy obtained if all the mislabeled data were removed. In addition, the vertical black dashed line represents the minimum number of iterations required to delete all the mislabeled data. Each curve color represents a proximity measure: in blue $Binary$, in red $NbEdges$ and in yellow $MaxAncestor$.

The proposed iterative framework improves the classification performances for both 20 % and 40 % noise levels. Although some small variations can be observed, the global trend shows how OA increases. The small variations may be explained by the fact that the n instances are removed at each iteration regardless of the class.

$Binary$ measure shows the highest improvement for 20 % noise level. $NbEdges$ shows the highest improvement for 40 % noise level. These results are corroborated with the curves displaying the number of removed outliers, where the $Binary$ measure shows a higher precision rate than the other measures especially at 20 % noise level. Considering the $MaxAncestor$ measure, the OA values drop after 100 iterations at 40 % noise level. In this case, the $MaxAncestor$ approach removes instances that belong to the same class. The same behavior can be observed for the $NbEdges$ measure at a 20 % noise level. The limitation of these approaches comes from the scaling step proposed by Breiman using equation (2), which does not lead to exactly comparable outlier scores between classes.

5. CONCLUSION

This work tackles the outlier detection problem. More precisely, the quality assessment of the outlier detection methods that use RF structure is shown. The obtained results show that the similarity measures computed by using the RF tree structure seem to improve traditional distances for the mislabeled data detection.

In addition, a new iterative learning framework using outlier detection strategies is proposed. The results show a strong potential for the improvement of land cover mapping techniques. For example, the proposed framework can allow the use of the past land cover maps to classify recent SITS. The instances whose labels change will be ideally removed from the training set.

The next research step is to refine this iterative learning strategy by firstly investigating the setting of the parameter n. In the

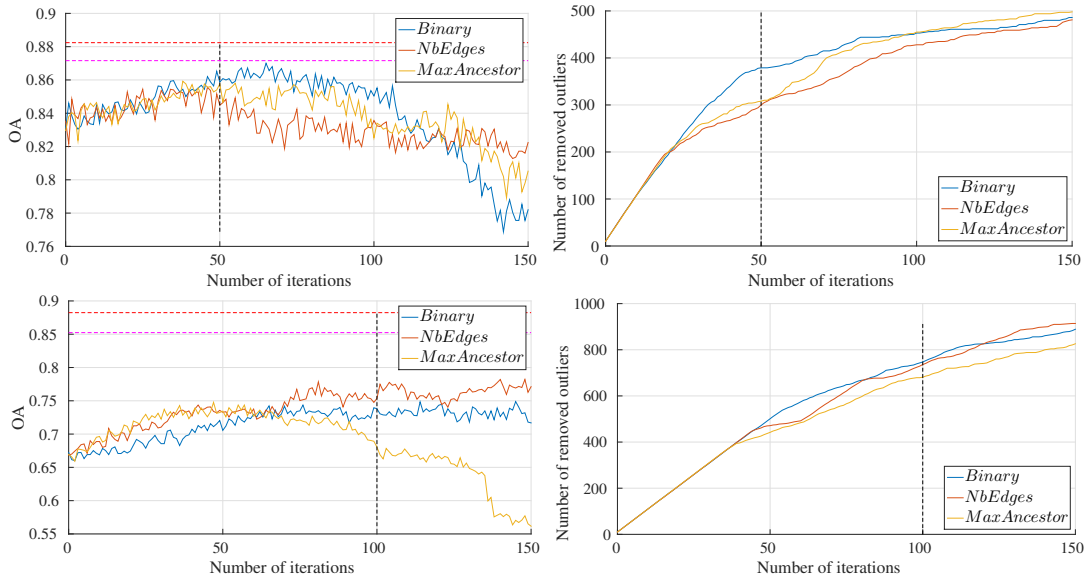

Fig. 3. Evaluation of the learning strategy. First column: OA values as a function of the iterations. Second column: the cumulative number of true detected outliers at each iteration. First row: 20 % noise level. Second row: 40 % noise level.

presented experiment, a small value for n is selected, increasing the number of iterations required to remove all the outliers. The trade-off between the accuracy and the number of required iterations should be further studied by testing the sensibility of the parameter n. In addition, a stopping criteria should be defined. Specifically, the use of the Out Of Bag (OOB) error of RF is being investigated.

6. REFERENCES

[1] C. Pelletier, S. Valero, J. Inglada, N. Champion, and G. Dedieu, "Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas," *Remote Sensing of Environment*, vol. 187, pp. 156–168, 2016.

[2] G. M. Foody, "The effect of mis-labeled training data on the accuracy of supervised image classification by SVM," in *IEEE International Geoscience and Remote Sensing Symposium 2015 (IGARSS)*, 2015, pp. 4987–4990.

[3] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 3, pp. 1014–1031, 2011.

[4] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 606–617, 2011.

[5] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-2, no. 3, pp. 408–421, 1972.

[6] C. E. Brodley, M. A. Friedl, et al., "Identifying and eliminating mislabeled training instances," in *American Association for Artificial Intelligence (AAAI) / Innovative Applications of Artificial Intelligence (IAAI)*, 1996, pp. 799–805.

[7] C-M. Teng, "Correcting noisy data," in *International Conference on Machine Learning*. Citeseer, 1999, pp. 239–248.

[8] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *ACM SIGMOD Record*, 2000, vol. 29, pp. 427–438.

[9] F. T. Liu, K. M. Ting, and Z-H. Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 1, pp. 3, 2012.

[10] M. M. Breunig, H-P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *ACM SIGMOD Record*, 2000, vol. 29, pp. 93–104.

[11] Franois Petitjean, Jordi Inglada, and Pierre Gancarski, "Satellite image time series analysis under time warping," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 8, pp. 3081–3095, Aug. 2012.

[12] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] A. Liaw and M. Wiener, "Classification and regression by Random Forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[14] C. Pelletier, S. Valero, J. Inglada, N. Champion, C. Marais Sicre, and G. Dedieu, "Effect of training class label noise on classification performances for land cover mapping with satellite image time series," *Remote Sensing*, In review.

[15] I. Tomek, "An experiment with the Edited Nearest-Neighbor rule," *IEEE Transactions on systems, Man, and Cybernetics*, , no. 6, pp. 448–452, 1976.

[16] G. O. Campos, A. Zimek, J. Sander, R. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle, "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study," *Data Mining and Knowledge Discovery*, pp. 1–37, 2015.

[17] B. Sluban, D. Gamberger, and N. Lavrač, "Ensemble-based noise detection: noise ranking and visual performance evaluation," *Data Mining and Knowledge Discovery*, vol. 28, no. 2, pp. 265–303, 2014.