5th International Workshop on Climate Informatics
September 24-25, 2015
Hosted by the National Center for Atmospheric Research in Boulder, CO

MONITORING TROPICAL FOREST FIRES USING PREDICTIVE MODELING FOR RARE CLASSES

Varun Mithal[1], Guruprasad Nayak [1], Ankush Khandelwal[1], Vipin Kumar[1], Ramakrishna Nemani[2], Nikunj Oza[2]

Abstract—**This paper presents a new predictive modeling framework designed to learn classification models from imperfectly labeled samples, in the absence of expert-annotated training samples, for identifying rare classes. Our results show that, under some reasonable assumptions, the classifiers trained from imperfectly labeled training data using this approach have performance comparable to the models trained using expert-annotated training data. This capability of learning from imperfect supervision is advantageous in a wide range of applications where the target class of interest is relatively rare and obtaining a precise labeling of even a small number of training samples is infeasible. We present the application of the framework for creating historical maps of forest fires from satellite data for the tropical forests. This new forest fire product identifies approximately 1 million sq. km. of burned areas in the tropical forests in South America and South-east Asia during years 2001-2014, which is more than three times of the total burned area reported by the state-of-art NASA products in these regions. We show validation of these results using burn-scars visible in satellite images to confirm the veracity of these forest fires.**

I. MOTIVATION

The *traditional* classification approach uses labeled training data to select the best classification model from a family of models. Since collecting labeled samples is often tedious and sometimes even infeasible, recent research in machine learning [1, 2, 3] has focused on developing algorithms to train classification models in scarcity of labeled training samples. In contrast, the focus of this paper is to address problem settings where acquiring even a small number of expert-annotated gold standard labeled samples for supervision is infeasible.

The problem of learning in the complete absence of labels is relevant in many important domains. Consider the task of mapping forest fires globally from satellite data in which the relationship between the explanatory variables and target variable varies in space and time. This necessitates learning separate customized classification models for each homogeneous data partition because learning a single global model, or applying a model trained in one partition on another heterogeneous partition leads to suboptimal performance. Since expert labeled data is not available in most parts of the world, there is a need to train classifiers in absence of labeled training data. The problem of mapping forest fires from satellite data is also challenging due to the ultra-skew class distribution. Traditional training algorithms, eg. logistic regression, are designed to learn the model parameters to achieve highest classification accuracy when the conditional probability ($P(y = 1|x)$) is thresholded at 0.5. For rare class problem settings, a more appropriate objective is to maximize the recall and precision of the rare class. Varying the threshold used on $P(y = 1|x)$ changes the precision and recall; increasing the threshold value decreases the recall while decreasing the threshold value decreases the precision. Incorrect choice of threshold can lead to very poor performance (because either recall or precision is too low) and thus it is critical to find the threshold that maximizes a desired combination of precision and recall.

Corresponding author: V Mithal, University of Minnesota, Twin Cities MN mithal@cs.umn.edu. [1] University of Minnesota [2] NASA AMES

II. METHOD

In absence of expert provided labels, we propose a new framework RAPT (RAre class Prediction in absence of True labels), which makes use of imperfect labels to train classifiers. In previous research it has been shown that if the imperfect labels are conditionally independent of the explanatory variables, the conditional probability from model trained using expert labeled training samples and the conditional probability from model trained using imperfectly labeled training samples give identical rankings for test instances. In case of expert labeled training samples, a threshold of 0.5 on the conditional probability is known to maximize accuracy. However, if the imperfect labels are perturbations of the true labels, then the threshold corresponding to maximum classification accuracy is not 0.5 and needs to be identified during training phase (Natarajan et al. [4]).

Moreover, for highly imbalanced class distributions, classification accuracy is not an appropriate metric, and it is desirable to *jointly maximize* the precision and recall of the rare class. Therefore, we select the threshold on conditional probability that maximizes the product of precision and recall. This is challenging as it requires the knowledge of the true labels of the validation set in order to accurately compute the precision and recall of the predictions corresponding to different candidate threshold values. We propose a method to estimate the geometric mean of precision and recall (*G-measure*) using only imperfect labels in context of rare class problems. We also prove that under the assumption of conditional independence between the imperfect labels and explanatory variables, the predictions from this classification model trained only using imperfect labels are comparable to the predictions from the classification model trained using expert provided labels.

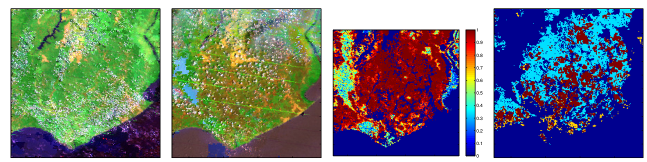
In RAPT framework, we use a combination step to aggregate the predictions from two conditionally independent views and significantly improve precision, while only reducing recall moderately in the context of rare class detection. More specifically, we use the imperfect labels as one of the views and predictions from the classification model as the other view and assign the instances predicted as positive in both views to positive class and remaining instances to the negative class.

Finally, we use a collective classification method [5] to improve the recall of the rare class by leveraging the guilt-by-association principle. The algorithm starts with the set of highly confident rare class test instances predicted by the combination step and iteratively increases coverage of the rare class by probing instances connected to the current set of confident positives.

III. EVALUATION

The RAPT framework was applied to create historical maps of forest fires from satellite data for the tropical forests. This new forest fire product identified approximately 1 million sq. km. of burned areas in the tropical forests in South America and South-east Asia during years 2001-2014, which is more than three times of the total burned area reported by the state-of-art NASA product (MCD64) in these regions.

Figure 1 shows the illustration of our approach for a 100 km. by 100 km. region in Indonesia, which experienced severe fire activity in the year 2006. Figure 1a and 1b show the multispectral composite before and after the fire event respectively. The burn-scar is clearly visible in the after-event composite. Figure 1c shows the burn-scar probability according to the classifier trained using RAPT. There is a good agreement between the $P(y = 1|x)$ and the burn scar visible in the Figure 1b. Figure 1d shows the comparison with respect to the state-of-art NASA burned area product MCD64. The pixels colored in *light blue* are detected only by the new scheme, while the pixels in *red* are detected by RAPT as well as MCD64. We verify the veracity of



(a) Before (b) After (c) $P(y = 1|x)$ (d) Comparison

Fig. 1: Illustration of burned area detection in a 100 km by 100 km area in Kalimantan, Indonesia.

the additional burned areas detect by RAPT (*light blue*) using manual inspection of burn-scars visible in spectral composites. Also, we computed difference in Normalized Burn Ratio (dNBR), which is an independently developed burn index, and show that the additional burned areas detected using RAPT have a similar distribution as the pixels detected only by MCD64. All these difference sources of evidence confirm that the additional burned areas detected by RAPT are indeed true burns that were missed by the state-of-art burned area products.

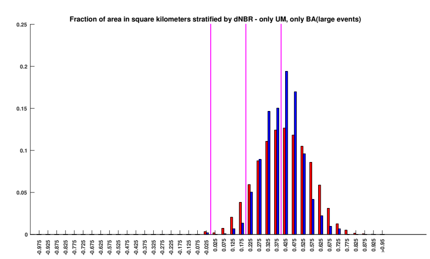


Fig. 2: dNBR distribution of only RAPT (*red*) and only MCD45 (*blue*) burned areas.

Acknowledgments This research was supported by the University of Minnesota Doctoral Dissertation Fellowship, National Science Foundation under Grants IIS-1029711 and IIS-0905581, and NASA grant NNX09AL60G. Access to computing facilities was provided by the University of Minnesota Supercomputing Institute.

REFERENCES

[1] O. Chapelle, B. Schölkopf, A. Zien, *et al.*, *Semi-supervised learning*, vol. 2. MIT press Cambridge, 2006.

[2] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pp. 109–117, ACM, 2004.

[3] B. Settles, "Active learning literature survey," *Technical Report University of Wisconsin, Madison*, vol. 52, pp. 55–66, 2010.

[4] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Advances in Neural Information Processing Systems*, pp. 1196–1204, 2013.

[5] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI magazine*, vol. 29, no. 3, p. 93, 2008.