9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

SPATIO-TEMPORAL CLASSIFICATION AT MULTIPLE RESOLUTIONS USING MULTI-VIEW REGULARIZATION

Guruprasad Nayak[1*] , Rahul Ghosh[1*], Varun Mithal[2], Xiaowei Jia[1], Vipin Kumar[1]

***Abstract*—In this work, we present a multi-view framework to classify spatio-temporal phenomena at multiple resolutions. This approach utilizes the complementarity of features across different resolutions and improves the corresponding models by enforcing consistency of their predictions on unlabeled data. Unlike traditional multi-view learning problems, the key challenge in our case is that there is a many-to-one correspondence between instances across different resolutions, which needs to be explicitly modeled. Experiments on the real-world application of mapping urban areas using spatial raster data-sets from satellite observations show the benefits of the proposed multi-view framework.**

I. MOTIVATION

Managing urban areas has become one of the most important developmental challenges of the 21st century[1]. Urbanization is beneficial from a sustainability standpoint since it is resourcefully more economical and environmentally less damaging to provide for a concentrated population than a dispersed one [1]. However, rapid unchecked urbanization can have adverse environmental and ecological issues [2], [3] along with poor living conditions [4]. Thus, an urban planning agenda is critical to ensure our sustainable living on the planet. This requires that we have regularly updated maps for urban land and population density to track their dynamics.

Raster images of reflectance data collected by satellites orbiting the Earth can be used for generating such urban maps. Raster data [5] are inherently spatio-temporal in nature where the underlying spatio-temporal field is observed at fixed locations in space and fixed points in time. Examples of spatio-temporal raster data-sets in environmental applications include time series of air quality measurements through ground sensors at a weather station, raster images of reflectance

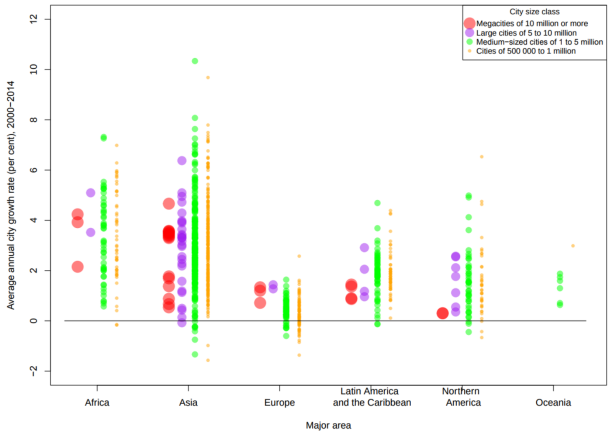

Fig. 1: Most urban expansion is occurring in developing nations where deploying resource-intensive methods is not an option. Figure courtesy of [1].

data collected by satellites orbiting the Earth. A key characteristic of raster data-sets is the resolution at which the data is collected. The same spatio-temporal field might be observed at different resolutions in space and time across different data-sets. Data-sets collected at different resolutions might differ in terms of the kind of features that they have and the availability of training samples at that resolution. However, since they all observe the same underlying spatio-temporal field, inferences made using data observed at different resolutions have to be consistent and thus, they can complement each other.

As an example, table I shows the spatial and temporal resolutions for few of the publicly available raster data-sets in the remote sensing domain. These data-sets are used for various environmental applications, prominently among which is the task of mapping land cover and land use across the globe over time [6]. From the table, one can clearly see that there is a wide range of available resolutions. While we may want high resolution maps of land use, using coarser resolution data-sets in conjunction can be very advantageous. Coarser spatial resolution satellite products such as MODIS (500m*500m pixels) are observed more frequently in

*These authors contributed equally to this article [1]University of Minnesota, {nayak013, ghosh128, jiaxx221, kumar001}@umn.edu
[2]LinkedIn, varunmithal@gmail.com

TABLE I: The variability in the available raster data-sets for the remote sensing domain

DATASET	TEMPORAL RESOLUTION	SPATIAL RESOLUTION	AVAILABILITY
MODIS	8 days	500 m	via NASA since 1999
LANDSAT	16 days	30 m	via USGS since 1972
SENTINEL	10 days	10 m	via ESA since 2014
PLANET-SCOPE	select dates	3 m	commercially available since 2016
WorldView	select dates	30 cm	commercially available since 2014

time (every 8 days), while very fine resolution products such as WorldView-3 (30cm*30cm pixels) might only be available for select dates. Coarser spatial resolution products have been collected for several decades, unlike some of the fine resolution products, so historical analyses are not possible with fine resolution data-sets. Moreover, it might be easier to procure training data at coarser resolutions since there are many good quality land cover products that already exist at this resolution. Finally, since every satellite product typically observes the Earth across different ranges of wavelengths, the features for the same location across different satellite products are different and hence complementary.

In this paper, we consider the problem of predicting a class of interest on a spatio-temporal field using raster data-sets at multiple resolutions. Generally, one operates in a setting where we are trying to predict the class of interest at a single resolution, with features and training samples provided for that resolution. In contrast, in our setting, training data is available for a limited number of instances at each resolution, that informs us of the presence or absence of the class of interest, when observed from that resolution. The different resolutions can be seen as different *views* that describe the same data set, on which a predictive model is being learned. The theory of multi-view learning, which considers predictive modeling settings with multiple *views* of the data, states that under the assumption that the feature spaces corresponding to the two views are independent of each other, it is possible for models trained separately on each *view* of the data to learn from the others. Therefore, in our case, instead of separately learning models to predict the phenomenon at different resolutions, we propose to simultaneously learn these models while maintaining the consistency of predictions between resolutions on the entire data set,

Fig. 2: This figure shows a small region in Minneapolis, USA and the corresponding labels (urban land) as observed from 30 m resolution and 10 m resolution. Note how only part of the region seems urban from the finer resolution although all of it is urban at the coarse resolution. Moreover, the fraction of urban fine resolution instances within a urban coarse resolution instance can have a large variability.

labeled or unlabeled. This has the following advantages (1) Utilizes an independent feature space (2) Utilizes unlabeled data and (3) Mitigates lack of labeled data from either resolution.

However, the key challenge in applying the multi-view learning approach in our case is the many-to-one nature of correspondence between instances from one view to the other. Figure 2 shows a caricature demonstrating this challenge in enforcing the consistency idea. The example considers the problem of predicting urban areas using two raster data-sets with different spatial resolutions. Note that the number of fine resolution instances classified as urban within an urban coarse resolution instance is variable within the data set. In this paper, we propose a multi-instance learning based strategy to handle this many-to-one correspondence between a pair of resolutions, coarse and fine, where the presence of even a single instance of the class of interest among the finer resolution instances makes the corresponding coarse resolution instance positive.

In a nutshell, we make the following contributions in this work

1) Formalize the problem of classifying phenomena across multiple resolutions in a multi-view framework.
2) Propose multi-instance learning strategy to implement the multi-view framework that handle the many-to-one correspondence across resolutions.
3) Demonstrate the utility of the proposed method on land cover mapping problems of great environmental significance.

II. METHOD

A. Problem Setting

In this paper, we consider the problem of classifying a spatio-temporal field on multiple resolutions. Data

instances at every resolution $1 \leq k \leq NR$ can be described through attributes $(\boldsymbol{x}, l, y)$ described as follows

1) Features $\boldsymbol{x} \in R^{D_k}$, where D_k is the dimensionality of the features observed in the kth resolution
2) Location l that describes the voxel id of the observation that encodes the location in space and the point in time where each instance is observed.
3) Label $y \in \{0, 1\}$ that encodes the presence ($y = 1$) or absence ($y = 0$) of the class of interest at location l

Goal: To learn classification models $\boldsymbol{w}_k$ at every resolution k that can take the features $\boldsymbol{x^k}$ observed for an instance on that resolution and predict its corresponding label y^k.

Training data: During the training phase, for each resolution k, we are provided a set T_l^k of $N_l^k > 0$ labeled samples $\{\boldsymbol{x}_i^k, l_i^k, y_i^k\}_{i=1}^{N_l^k}$. In addition, we have unlabeled data for a large region RU of the spatio-temporal field at every resolution i.e we have features $\boldsymbol{x}_i^k$ for every observable location l within the region RU at resolution k, forming a unlabeled training data set T_u^k of N_u^k samples $\{\boldsymbol{x}_i^k, l_i^k\}_{i=1}^{N_u^k}$.

B. Multi-view framework

Classification models $\boldsymbol{w}_k$ are learned at every resolution $k \in \{1 \cdots, NR\}$. In particular, classifier $f_k(\boldsymbol{x^k}; \boldsymbol{w}_k)$ at resolution k with parameters $\boldsymbol{w}_k$ models $Pr(y^k = 1|\boldsymbol{x}^k)$, where y^k is the label at that resolution. We place no restriction on the actual form of f. It could take any form like LSTMs for temporal data or CNNs for spatial data or more traditional classifiers like a neural network with one hidden layer. However, instead of learning the classifiers at different resolutions independently, as would be the case in a conventional approach, we propose to use the large number of unlabeled data available on the same region RU of the spatio-temporal field to enforce consistency in predictions across resolutions and thus, make the models on different resolutions learn from each other. Thus, the objective function takes the following form,

$$O(\boldsymbol{w_1}, \cdots, \boldsymbol{w_{NR}}) = \sum_{k=1}^{NR} L(T_l^k; \boldsymbol{w_k}) + \sum_{k_1=1}^{NR-1} \lambda_{k_1} \sum_{k_2=k_1+1}^{NR} D(pred(T_u^{k_1}; \boldsymbol{w}_{k_1}), pred(T_u^{k_2}; \boldsymbol{w}_{k_2})) \tag{1}$$

The first term in the objective function is the loss over labeled samples while the second term is a regularization term that enforces the consistency of predictions across resolutions on the unlabeled instances i.e $L(T_l^k; \boldsymbol{w_k})$ is the loss over the labeled training instances T_l^k on the kth resolution and the function $D()$ captures the consistency of the predictions between every pair of resolutions.

1) Defining consistency across resolutions: The choice of function $L()$ in equation 1 is standard. Defining the consistency function $D()$ is non-trivial because of the lack of an one-to-one mapping between instances across a pair of resolutions, as would be the case in traditional multi-view problems. Given a pair of resolutions - coarse and fine, one can define a many-to-one mapping of instances from the fine resolution to the coarse resolution by using a nearest neighbor approach. i.e every fine resolution instance is assigned to the coarse resolution instance with the closest location to it. Subsequently, the consistency of predictions on the unlabeled instances between a pair of resolutions boils down to defining the consistency between every coarse resolution instance and its corresponding fine resolution instances. In particular, the consistency term in equation 1 can be rewritten as,

$$D(pred(T_u^{k_1}; \boldsymbol{w}_{k_1}), pred(T_u^{k_2}; \boldsymbol{w}_{k_2})) = \sum_{i \in T_u^{k_1}} d(\boldsymbol{x_i^{k_1}}, \{\boldsymbol{x_j^{k_2}} | j \in S_i \text{ and } j \in T_u^{k_2}\}, \boldsymbol{w_{k_1}}, \boldsymbol{w_{k_2}}) \tag{2}$$

where the summation is over all unlabeled instances i in the coarser resolution k_1. Also, the set S_i denotes the unlabeled instances in fine resolution k_2 that are closest in location to instance i from the coarser resolution.

C. Multiple Instance Learning (MIL) solution

Multiple Instance Learning (MIL) [7], [8] considers the problem of learning predictive models to label groups of instances, in contrast to traditional settings where the goal is to label individual instances. Typical MIL classification settings operate under the presence-based assumption[7] which states that a group has a positive label when at least one of its constituting instances has a positive label and it has a negative label when all of its constituting instances have a negative label as well. MIL forms a direct way to define function $d()$ in equation 2 that models the many-to-one relationship between an instance of a coarse resolution and its corresponding instances in the finer resolution. Given an instance i from a coarse resolution k_1 and its corresponding instances S_i from a fine resolution k_2, the prediction for the coarse resolution label can be written in two ways - first using the model on

TABLE II: Comparison with baselines: balanced data-sets (Accuracy)

METHODS		DATASETS	
		ROME	MINNEAPOLIS
LANDSAT	LogReg	0.877	0.887
	Conc Features	0.822	0.821
	Semi Supervised	0.876	0.888
	Multi-Res	**0.913**	**0.917**
SENTINEL	LogReg	0.852	0.884
	Conc Features	0.874	0.896
	Semi Supervised	0.884	0.899
	Multi-Res	**0.944**	**0.952**

resolution k_1 as $Pr(y_i^{k_1}|\boldsymbol{x_i^{k_1}}) = f_{k_1}(\boldsymbol{x_i^{k_1}}; \boldsymbol{w}_{k_1})$. Secondly, the label at k_1 can also be predicted using corresponding instances on k_2 using the MIL assumption as $Pr(y_i^{k_1}|\{\boldsymbol{x_j^{k_2}}|j \in S_i\}) = \max_{j \in S_i} f_{k_2}(\boldsymbol{x_j^{k_2}}; \boldsymbol{w}_{k_2})$. Note that taking the maximum of the probabilities for instance-level for constituting instances is one way to implement the presence-based MIL assumption[9]. Thus, the function $d()$ in equation 2 can be defined as,

$$d(\boldsymbol{x_i^{k_1}}, \{\boldsymbol{x_j^{k_2}}|j \in S_i \text{ and } j \in T_u^{k_2}\}, \boldsymbol{w_{k_1}}, \boldsymbol{w_{k_2}}) = \left(f_{k_1}(\boldsymbol{x_i^{k_1}}; \boldsymbol{w}_{k_1}) - \max_{j \in S_i} f_{k_2}(\boldsymbol{x_j^{k_2}}; \boldsymbol{w}_{k_2})\right)^2$$

Since we use gradient descent algorithm to learn the optimal parameters for our models, we want our objective functions to be differentiable and hence, in our implementation, the max function is replaced by its differentiable softmax approximation.

III. EVALUATION

The methods proposed in this paper are evaluated on data-sets of 2 different regions namely Minneapolis and Rome, from a real world application - urban area detection, that use satellite-collected observations of different locations to automatically track changes on the surface of the Earth. We use data-sets from two satellites available at two different resolutions. Landsat 8 satellite data product having 30m spatial resolution(3660×3660 pixels) and Sentinel 2A data product having 10m spatial resolution(10980×10980 pixels). These two sources act as the coarse and fine resolution respectively. In each data-set we take 200 training samples and 60000 test samples on each resolution. In addition to those we also have 10000 unlabeled coarse-resolution and correspondingly, 90000 unlabeled fine-resolution samples (there are 9 Sentinel pixels for each Landsat pixel). We compare the following methods on the above data-sets:

1. Coarse-Resolution Logistic Regression and Fine-resolution Logistic Regression Separate models are trained on both resolutions using their respective labeled samples.

2. Concatenated Features Logistic Regression This method uses a concatenated feature space using features from the other resolution. For each coarse-resolution pixel, the corresponding fine-resolution pixels are averaged and concatenated to have an extended feature-set. The same is done for each fine-resolution pixel. Separate models are learned at each resolution using the extended feature sets.

3. Semi-Supervised Logistic Regression This is an extension of baseline 1 where in addition to labeled samples at a given resolution, unlabeled samples are also used to regularize predictions at that resolution. Note that models at different resolutions are still trained separately.

4. Multi-Resolution Logistic Regression Proposed algorithm which uses the limited labeled data available in the respective resolutions and enforces consistency in predictions on unlabeled data between resolutions.

A. Results

Table II reports the accuracy for the two data sets for the urban mapping application i.e Rome and Minneapolis. First, we have *LogReg*, that is trained with just labeled samples from each resolution. If we concatenate features from the other resolution (*Conc Features*), we observe that the performance goes down since there are not enough samples to handle the increased dimentionality. Using unlabeled data in a semi-supervised fashion yields some benefit, especially in the finer resolution. However, the most gains are found in using unlabeled data to enforce consistency between resolutions as it brings out the benefit of an independent feature space to learn from.

IV. CONCLUSION

In this paper, we formalized the problem of classifying spatio-temporal phenomena simultaneously at multiple resolutions in a multi-view framework. The multi-view framework helps to regularize the models trained on individual resolutions by enforcing consistency of predictions across resolutions on the large number of freely-available unlabeled data. Unlike traditional multi-view learning scenarios, the multi-resolution classification task involves a many-to-one correspondence between views of the data, which the proposed methods in the paper learn explicitly through multiple instance learning and attention mechanism. Experiments on urban mapping data-sets show the utility of utilizing unlabeled data through the multi-view framework. Future work involves investigating other approaches to modeling the consistency across resolutions such as through the attention mechanism [10] that has shown much promise in domains such as Natural Language Processing (NLP).

REFERENCES

[1] U. N. D. of Economic and S. A. P. Division, "World urbanization prospects: The 2014 revision," vol. ST/ESA/SER.A/352, 2014.

[2] X. Li and P. Gong, "Urban growth models: progress and perspective," *Science bulletin*, vol. 61, pp. 1637–1650, 2016.

[3] N. Ranger, S. Hallegatte, S. Bhattacharya, M. Bachu, S. Priya, K. Dhore, F. Rafique, P. Mathur, N. Naville, F. Henriet, and C. Herweijer, "An assessment of the potential impact of climate change on flood risk in mumbai," *Climatic change*, vol. 104, no. 1, pp. 139–167, 2011.

[4] *https://www.theguardian.com/cities/2014/nov/24/mumbai-verge-imploding-polluted-megacity*.

[5] G. Atluri, A. Karpatne, and V. Kumar, "Spatio-temporal data mining: A survey of problems and methods," *ACM Computing Surveys (CSUR)*, vol. 51.4, p. 83, 2018.

[6] A. Karpatne and et al, "Monitoring land-cover changes: A machine-learning perspective," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4.2, pp. 8–21, 2016.

[7] J. Foulds and F. Eibe, "A review of multi-instance learning assumptions," *The Knowledge Engineering Review*, vol. 25, no. 1, pp. 1–25, 2010.

[8] V. Cheplygina, D. Tax, and M. Loog, "On classification with bags, groups and sets," *Pattern recognition letters*, vol. 59, pp. 11–17, 2015.

[9] S. Ray and M. Craven, "Supervised versus multiple instance learning: An empirical comparison," *Proceedings of the 22nd international conference on Machine learning*, 2005.

[10] A. Galassi, M. Lippi, and Torroni, "Attention, please! a critical review of neural attention models in natural language processing," *arXiv preprint*, 2019.