Exploiting Data and Human Knowledge for Predicting Wildlife Poaching

Swaminathan Gurumurthy
Carnegie Mellon University
sgurumur@andrew.cmu.edu

Lantao Yu
Shanghai Jiao Tong University
yulantao@apex.sjtu.edu.cn

Chenyan Zhang
Southeast University
cy.zhang@seu.edu.cn

Yongchao Jin
World Wild Fund for Nature,
China
jinyongchao@orientscape.com

Weiping Li
World Wild Fund for Nature,
China
wpli@wwfchina.org

Xiaodong Zhang
Huang Ni He Forest Bureau
1009595567@qq.com

Fei Fang
Carnegie Mellon University
feifang@cmu.edu

ABSTRACT

Poaching continues to be a significant threat to the conservation of wildlife and the associated ecosystem. Estimating and predicting where the poachers have committed or would commit crimes is essential to more effectively allocate patrolling resources. The real-world data in this domain is often sparse, noisy and incomplete, consisting of a small number of positive data (poaching signs), a large number of negative data with label uncertainty, and an even larger number of unlabeled data. Fortunately, domain experts such as rangers can provide complementary information about poaching activity patterns. However, this kind of prior knowledge has rarely been used in previous approaches.

In this paper, we contribute solutions to the predictive analysis of poaching patterns based on very limited data and human knowledge. We propose to elicit quantitative information from domain experts through clustering data points based on geographical features and asking domain experts to provide an estimate for each cluster. In addition, we propose algorithms that exploit qualitative and quantitative information provided by the domain experts to augment the dataset and improve learning. In collaboration with World Wild Fund for Nature, we show that incorporating human knowledge leads to better predictions in a conservation area in Northeastern China where the charismatic species is Siberian Tiger. The results show the importance of exploiting human knowledge when learning from limited data.

ACM Classification Keywords

I.2.6 Artificial Intelligence: Learning; K.4.m Computers and Society: Miscellaneous; H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Computational Sustainability; Data Science; Machine Learning; Wildlife Conservation.

INTRODUCTION

Poaching is a profitable crime that threatens endangered species. Trade of ivory, tiger bones and skins are well-known examples of some of the incentives for this crime [2, 15]. Wildlife conservation agencies aim to protect the wildlife and their habitats from increased poaching and illegal trade. To achieve this goal, they send rangers to patrol in protected conservation areas [11]. However, because of limited patrolling resources, it is impossible to monitor all intrusion routes and protect the entire area. Therefore, it is critical to understand how poachers make decisions, predict the patterns of poaching to strategically allocate patrolling resources in order to detect and deter poaching activities, and reduce potential harm to wildlife and their habitats.

Rangers record their findings, including animal signs and poaching activity signs, (e.g., snares placed by poachers) during the patrol, and therefore one can analyze these records to get insights of the poaching patterns. There have been several previous projects that provided predictive tools through designing machine learning models trained and evaluated using real-world data from two conservation sites in Uganda [13, 8, 5, 6]. In these works, a dataset is created based on past patrols and geographical information. The conservation area is discretized into grid, and each data point corresponds to a grid cell during a particular time period, with the label being whether or not there were poaching activities in the cell in that time period. However, due to the limited patrolling resources

and the inability of the rangers to detect all poaching activities signs during the patrols, the dataset often consists of a small number of positive data (cells with poaching activities found by rangers), a large number of negative data with label uncertainty (cells patrolled but with no poaching activities found), and an even larger number of unlabeled data (cell not patrolled). Despite the effort made towards addressing these challenges, sparsity of positive data is still a major challenge and the previous works failed to (i) exploit the unlabeled data to assist the learning; (ii) exploit human knowledge from domain experts, such as conservation site managers and rangers in a quantitative way.

The rangers have extensive experience in the field and have been interacting with poachers for years. They typically choose patrol routes with the goal of detecting and deterring poaching activities, and therefore the unlabeled data implicitly encodes their understanding about where the poaching activities may not take place. In addition, their knowledge about the poachers' capabilities, tactics, and the area goes beyond what is recorded in the dataset. The information they can provide in order to build our understanding of the adversaries is critical. The key challenges here are to elicit quantitative information from them and to make efficient use of the information.

In this paper, we focus on exploiting both the data and human knowledge to predict poaching activities and provide the following contributions: (1) We present a questionnaire design to elicit quantitative information about poaching threat from domain experts. We use k-means clustering to divide the conservation area into small regions or clusters, and present the clusters to the domain experts. In the questionnaire, we ask them to provide an estimate of the poaching threat for each cluster. (2) We provide two approaches of exploiting human knowledge to enhance machine-learning-based predictive analysis. The first approach is to sample negative data points from unlabeled data. The second approach is to sample positive and/or negative data points from cluster-based estimation of poaching threat. (3) We evaluate our proposed approaches using data from 2014-2018. Our experimental results show that incorporating human knowledge can improve the performance of the predictive model.

While we use anti-poaching as the domain of focus, the methodologies introduced in this paper are general and applicable to a large number of security problems with limited real-world data and experienced human experts, such as predicting crimes in areas with insufficient record.

RELATED WORK
There has been some previous work focusing on understanding and predicting poaching activities. [12] identifies that areas with certain features are at high risk of poaching and demonstrates that increasing the number of patrols will effectively reduce poaching activities using data from Nyungwe National Park. Similar work has been done at Tsavo National Park, [17] performs geospatial analysis on the physical environment to find correlations between geographical features and poaching incidents. More recently, machine learning based approaches have been explored to predict poaching. Based on real-world

data from Queen Elizabeth Protected Area, [13] uses a Dynamic Bayesian Network that explicitly models the dependencies between occurrence and detection of poaching activities, as well as the temporal pattern of poaching. [8] designs an ensemble of decision trees which incorporate spatial correlation of poaching to account for the undetected instances. [5] provides a hybrid model that combines decision trees and Markov Random Fields [5] to exploit the spatio-temporal correlation of poaching activities. [6] proposes to weigh the negative data points in the training set based on patrol effort so as to account for the label uncertainty. However, the challenge of having limited data is not fully resolved and human knowledge is only used in very limited ways in previous works such as to select which features to be considered and to represent the dependency and correlation relationships.

In addition to wildlife poaching, predicting other types of crimes based on real-world data has been studied using general principles such as "crime predicts crime" in criminology. Machine learning models and algorithms including Bayesian networks, k-nearest neighbors, neural networks, decision trees, and support vector machines [18, 7] have been applied. However, most of the work rely on a sufficiently large dataset and ignore the fact that there are undetected or unreported crime instances. In many domains including wildlife poaching, real world data that can be used to build a machine learning model for predicting crime pattern is very limited and sometimes incomplete. Learning from such datasets and making prediction is very difficult.

Efforts have been made to study learning with limited labeled data and incorporating domain knowledge in the learning phase. For example, [14] provides empirical studies to examine several aspects including independence or dependence amongst features, the amount of unlabeled data versus amount of labeled data, and the effect of label noise. [21] proposes a guideline to incorporate domain knowledge into inductive machine learning while pointing out the difficulty in finding a universal solution, given the variety and diversity of domain knowledge. Progress has been made to address the problem. [22] introduces an approach to learn an instance-based classification model from subgroups of instances and a human estimation of the proportion of instances with certain labels. A method to learn from uncertain query responses from domain experts is also proposed by [9]. [1] gives a method to integrate domain knowledge in the field of breast cancer risk assessment by adding virtual instances to the original dataset.

DOMAIN DESCRIPTION AND REAL-WORLD DATASET
Wildlife poaching is a global challenge and poaching activities differ in many ways, including the poachers' goals, target species, tools, tactics and capabilities. For example, in 2012, poachers armed with grenades and AK-47s killed more than 300 elephants in Bouba Ndjidah National Park, Cameroon [16]. In 2012-2013, 70 lions died after eating poisoned food prepared by poachers in southern Africa [20]. Among the different types of poaching activities, snare poaching is one of the most prevalent ways of poaching and is the main focus of this paper. Due to the low cost of making a snare (as low as $0.5-$2 per snare based on private conversations with experts from

WWF) and the low skill requirement, snare poaching crimes committed by organized poachers as well as local villagers threaten various species in many countries across continents.

Supported by government and non-government conservation agencies, rangers conduct patrols in order to detect snares before the wildlife gets trapped and deter the poachers. Rangers take records as they patrol, and in recent years, many conservation sites have started using software tools such as MIST and SMART to track and store the patrol routes and the records. However, the records only include poaching activities that are detected by the patrollers during their patrols. Some poaching activities were not recorded if the patroller did not patrol in the right area at the right time or did not find the well-hidden signs of poaching. Due to the limited patrolling resources and equipments, the records only cover a small portion of the poaching activities. In addition, wildlife animals are silent victims of poaching, making it hard to find other sources of reports on poaching activities.

Construction of Dataset
In this paper, we focus on Huang Ni He National Nature Reserve (HNHR), a conservation area spanning about 75 sq. km in Northeastern China where the charismatic species is Siberian Tiger. Snare poaching is the main poaching activity of concern in this area. The rangers do not go on patrols all year around, but instead mostly patrol in autumn and winter seasons as there is a high poaching threat in these seasons. Based on geographical data, historical patrol records in 2013-2017, and recent records of 2017-2018 winter season patrol in HNHR, we divide the area into 1km by 1km grids and construct a dataset where each data point corresponds to a grid cell in a patrol season. The label for each data point indicates whether or not there were any poaching activities in the grid cell in that patrol season. The features of each data point include: the distance from each area to the closest stream, village, patrol post, river, marsh, village road, provincial road, national road, highway, conservation boundary; land type, the elevation and slope of each area; patrol length in last patrol season. Thus, we use a total of 14 features, of which 10 are distance features.

Here we provide details about how the dataset is constructed:
(i) *Geographical Features*: Based on publicly available digital elevation model (DEM) data of the conservation area with 90m resolution [1], we derive the elevation and slope of the whole area through functions in Spatial Analyst Tools in ArcMap 10.2. We get the land type and detailed locations of road, river and villages in the format of shapefiles by digitalizing the forest map of HNHR manually. To get various distance features used in our model, we create a grid using MMQGIS tool in QGIS 2.14 and use NNjoin tool to compute the distance from the corresponding shapefiles of rivers, roads and villages. To ease the training of the model, we further normalize the geographical feature values into the range of 0 to 1 except for land type feature.
(ii) *Patrol Features*: Through the anti-poaching project supported by WWF, historical patrol routes as well as animal activities and poaching activities detected during patrols are

recorded in SMART [2]. Based on the exported file of historical waypoints, we compute the patrol length in each grid cell by splitting the patrol line into fixed intervals using *QChainage* tool and counting the intervals using Analysis Tools (*Count Points in Polygon* function) in QGIS 2.14. We also normalize the patrol length feature into the range of 0 to 1.
(iii) *labels*: To get the label of each data point, we count the detected poaching activities by importing the SMART observation data into QGIS and performing spatial joins using Data Management Tools (*Join Attributes by Location* function) in QGIS 2.14. The labels are then binarized by setting all areas with non-zero poaching activity detections to one (Positive Data Points) and the rest to zero (Negative Data Points).

In addition to historical data collected from 2013-2017, we also get recent records in 2017-2018 winter season in HNHR, which are collected through 34 patrol shifts spanning over 3 months of patrol. Each patrol shift took 10265.24 seconds (2.85 hours) on average.

Challenges in the Dataset
The lack of patrolling resources leads to a dataset with several challenges and peculiarities. It suffers from significant class imbalance, sparsity and noise in negative labels.

In fact, as can be seen in Figure 1-3, the dataset is extremely skewed with very few positive examples (29.5 on average[3]) in contrast to a large number of negative examples (9310.5 on average). In addition, only a small portion of the area is patrolled every year, leading to a very larger set of unlabeled data (44493.25 on average).

Detecting snares requires experience and in many cases, the snares cannot be found easily, given the tree branches in the background, especially when one is walking. Figure 4 shows an example snare that can't be easily detected by an unexperienced ranger. As a result, the patrollers might have simply missed the snares in certain regions while patrolling and hence a lot of the negative data points might indeed be positive. Therefore, we expect noise in the negative labels in our dataset.

METHODOLOGY
Given the dataset, we aim to train a model that can predict or estimate the label for any data point using geographical and patrol features. Due to the challenges of the dataset, training a machine learning model from it directly is very challenging. Hence we propose several approaches to elicit and exploit human knowledge. First, we collect explicit domain knowledge from experts and incorporate the domain knowledge in our model. Second, we use data duplication to alleviate the dataset skew towards negative examples. Third, most of the unlabeled data can be used as a proxy for negative data since the patrollers didn't even consider it worth visiting these areas. We will discuss these approaches in more detail in subsequent sections.

_____
[1]http://www.gscloud.cn/

_____
[2]http://smartconservationtools.org/
[3]As the patrols are still ongoing, the data in 2017-2018 is not complete and is excluded from the computation of average.
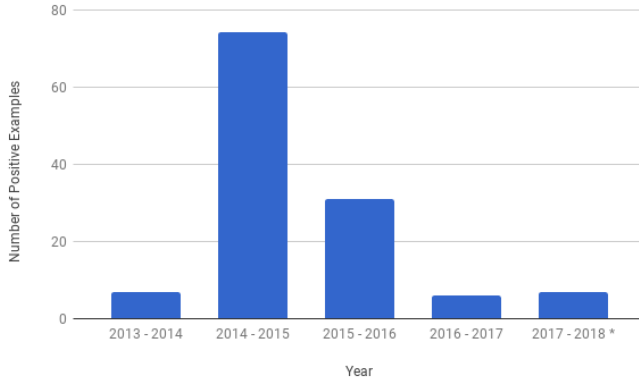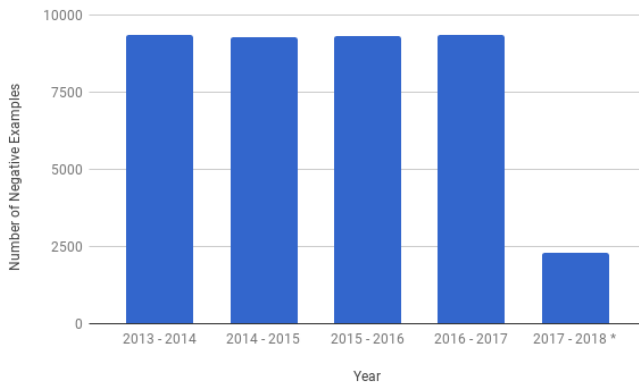
Figure 1. Number of Positive Examples

Figure 3. Number of Unlabeled Examples

Figure 2. Number of Negative Examples

Figure 4. A snare that is hard to detect.

Several different machine learning models have been proposed previously to predict poaching threats. However, this paper focuses on using expert knowledge to enhance machine learning based predictive tools. Therefore, we apply our approaches to two basic machine learning models to illustrate the importance of using expert knowledge: a bagging ensemble of decision trees and a neural network based model.

Eliciting Information from Domain Experts

Given the peculiarities of the dataset, it is very important to elicit some explicit or implicit information from domain experts so that we can make use of the unlabeled data as well as the noisy labeled data. It should be noted that several factors need to be kept in mind while collecting domain knowledge from the experts. Firstly, it is not possible for the experts to give very accurate and fine-grained information (e.g, the specific probabilities for every region). Second, the experts can't be expected to label huge amounts of data given the limited amount of resources and time they have. Third, it should be expected that the information provided by the experts can be noisy. Hence, we have to settle with a limited amount of information which is coarse grained and noisy. But at the same time we need to ensure that we are able to extract enough information to tackle the bottlenecks in our machine learning models. With these in mind, we develop the following method to elicit information from the domain experts.

Cluster Labels

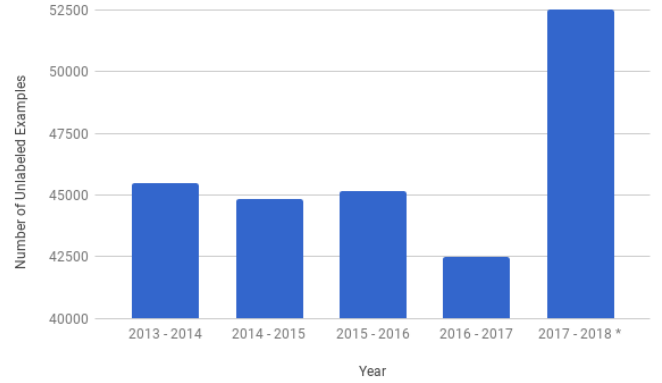Recognizing the fact that it is difficult for the experts to provide tentative labels/threat probabilities for every single region in the conservation area, and that fine grained labels provided by the experts can be very noisy, we propose a method to obtain weaker yet cleaner labels from the experts. Instead of asking the experts to score each individual region, we first group these regions into clusters using K-means clustering in the feature space. We then provide these clusters to the experts and ask them to provide a score for each cluster from 1 to 10, where 1 corresponds to minimum threat level and 10 corresponds to maximum threat level. To decide the number of clusters to be used in this procedure, we asked the domain experts for feedback. They believed that $40 - 50$ clusters would be a reasonable number to ensure consistency across the labeling on a given set of clusters. Following their advice, we repeat this procedure twice, first with 40 clusters and then with 50 clusters. Doing it twice helps account for any inconsistency in the allocated weak labels. This gives us two sets of clusters and their corresponding weak labels. This approach helps us extract useful information about the threat level without causing much cognitive burden on the experts. Specifically,

Figure 5. Visualization of the 40 clusters
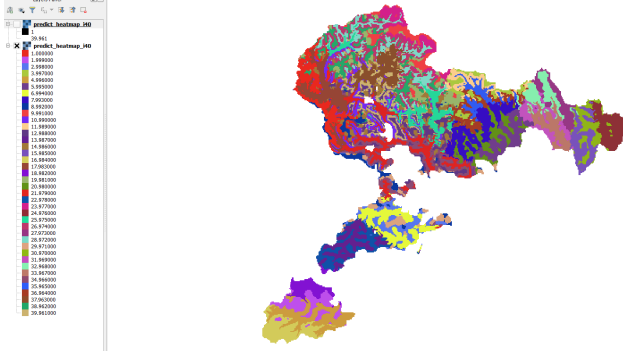
now the experts only need to provide 90 labels rather than 75000 labels. See Figure 5 for visualizations of the 40 clusters in the map.

Constructing Prior Knowledge

Although the prior knowledge collected from domain experts is very weak, it can be incorporated into the machine learning model in a variety of ways. For example, we sample the unlabeled examples with the highest confidence scores and add that to the positive examples. With these confidence scores, we can also design a weighting mechanism over the examples or use them to regularize the models. In order to do this we need to form a mathematical model to calculate these confidence scores for each example. We employ different models to calculate these confidence scores for different types of domain knowledge.

Cluster Labels

Here, we have direct access to the probability values $P(Y = 1|X)$. We have a score (from 1 to 10) for each cluster. The scores can then be directly converted to probability values by dividing the scores by 10. But we have these probability values for each cluster set individually. In order to combine these two set of cluster probabilities, we propose to use Bayes Rule. Specifically, given the cluster sets C^1 and C^2 and the corresponding probabilities $P(Y = 1|X \in C_i^1)$ and $P(Y = 1|X \in C_j^2)$ we want to estimate $P(Y = 1|X \in C_i^1, C_j^2)$.

$$P(Y = 1|X \in C_i^1, C_j^2) = \frac{P(X \in C_i^1, C_j^2|Y=1)P(Y=1)}{P(X \in C_i^1, C_j^2)}$$

We can assume that the $X \in C_i^1$ and $X \in C_j^2$ are independent events given Y. Hence the above equation can be decomposed into

$$P(Y = 1|X \in C_i^1, C_j^2) = \frac{P(X \in C_i^1|Y=1)P(X \in C_j^2|Y=1)P(Y=1)}{P(X \in C_i^1)P(X \in C_j^2)}$$

We can again apply Bayes Rule and get

$$P(Y = 1|X \in C_i^1, C_j^2)$$
$$= \frac{P(Y = 1|X \in C_i^1)P(X \in C_i^1)P(Y = 1|X \in C_j^2)P(X \in C_j^2)P(Y=1)}{P(Y=1)P(Y=1)P(X \in C_i^1)P(X \in C_j^2)}$$
$$= \frac{P(Y = 1|X \in C_i^1)P(Y = 1|X \in C_j^2)}{P(Y=1)}$$

Again, $P(Y = 1)$ is a constant, so the equation finally becomes:

$$P(Y = 1|X \in C_i^1, C_j^2) = \frac{P(Y = 1|X \in C_i^1)P(Y = 1|X \in C_j^2)}{Z} \tag{1}$$

We can now easily compute the threat confidence of any data point given the clusters it belongs to and the individual cluster scores. We can ignore the denominator while computing the confidence since it's just a constant.

Data Augmentation

As discussed earlier, the dataset has a lot of peculiarities that make the straightforward application of a machine learning algorithm extremely difficult. To be specific, there are very few positive examples compared to the large number of negative examples and even more unlabeled examples. In fact, most of the unlabeled examples would be negative. These issues can be mitigated to an extent using data augmentation. In this work, we try three ways for augmenting the data.

Data Duplication

Since, the number of positive examples are very low compared to the negative examples, we choose to duplicate the positive examples to balance the dataset. With this approach, we ensure that the number of positive examples and negative examples used for training the model remains the same level.

SMOTE

As an alternative to data duplication, we also tried another over-sampling technique to see if more complicated oversampling procedures can be useful. One of the most common alternatives to Data Duplication is called SMOTE: Synthetic Minority Over-sampling Technique[4]. In this technique, new examples are created by interpolating between a positive sample and a randomly chosen example from it's k-nearest neighbors. This procedure is repeated until the dataset becomes balanced.

Negative Sampling

As established previously we know that most of the unlabeled examples are low threat regions, since the experts chose not to explore those regions. Hence we add a random partition of the unlabeled set into the negative set during training.

Prior Knowledge for Positive Sampling

Using the prior knowledge extracted from human experts, we formulated a way to calculate a score and weak label for all data points. We use this score to obtain the most highly ranked points from the unlabeled set and add that to the positive set during training.

Model Implementations

Bagging Ensemble Decision Tree

We use bagging ensemble decision tree [3] with 1000 trees where each base tree is trained using only 10 percent of the total training data. We use entropy to compute the information gain at each node. We use this as our default classifier due to the superior performance (empirically demonstrated in the experiments). We use the implementation provided scikit-learn with the above mentioned parameters to train the model.

Neural Networks

We also use a three-layer feed forward neural network [10] with 8 neurons on the first layer , 4 neurons on the second layers and a single neuron in the last layer spitting out the threat probability for that data point. We use relu nonlinearity in the first and second layers and a sigmoid at the output. To predict the final output we use an ensemble of 100 such neural networks. We train using the Adam optimizer with learning rate = 0.001, beta1= 0.9 and beta2= 0.999. We keep a constant batch size of 64 throughout training. We also add an l2 penalty on the weights with weight decay constant of 0.1.

EVALUATION

In this section we evaluate our model and showcase the importance of each component of the model. We compare these models on 3 popular metrics. An earlier version of the model (presented as one of the baselines) was also deployed in a conservation area during the months of November to February. We also present the deployment results. The data collected from the deployment is also merged into the dataset while we report the results.

Metrics

We evaluate our model performance using 4 metrics precision, recall, F1 score and the ll score. We choose to report the ll score along with the precision, recall and F1 scores because it offers better discriminability since it's not bounded between $[0, 1]$. We don't include the more popular aggregate precision/recall score (AUC score here) because we observed that it was sensitive to false positives and gave a high score even when the number of false positives were high.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$F1score = \frac{2*Recall*Precision}{Recall + Precision}$$

$$llscore = \frac{Recall*TestSetSize}{TruePositives + FalsePositives}$$

Training and Testing

Given the limited number of positive samples in the dataset, it is infeasible to separate the dataset into a test set and a training set without getting a biased evaluation of our model. Hence, we used 4-fold cross validation to train and test our model performance. But we still observed very high variance over the cross validation runs. Thus, we repeat 4-fold cross validation multiple times and average the results to get a more accurate estimate of the scores. The dataset here includes the historical data collected by patrollers and the data collected during the deployment of an earlier version of our model (with only decision trees and negative sampling). cross-validation is performed on this merged dataset and the mean of precision, recall, ll score and F1 score are reported on each model after multiple evaluations.

Table 1. Scores for our model and contribution of each component

Models	LL score	Recall	Precision	F1 score
Random decisions	0.51	0.5	0.004	0.008
DT without DD	0.0	nan	0.0	0.000
DT with DD	14.60	0.31	0.17	0.219
DT with SMOTE	11.19	0.35	0.12	0.179
DT with DD, PS	4.99	0.35	0.05	0.087
DT with DD and NS	14.05	0.27	0.19	0.223
DT with DD, NS, PS	15.42	0.31	0.18	0.227
NN with DD, NS	2.47	0.48	0.02	0.038

Evaluation on Dataset

In this section we evaluate our model on the dataset and also show the impact of each component of the model. 1 contains the precision, recall, F1 and the ll scores for the model and multiple baselines. In the experiments listed in the Table 1, DD indicates Data Duplication (i.e, duplicate the positive data to match the number of negative examples). We find that data duplication is very crucial. The model completely fails (predicts negative labels for every example) if we remove this component. In fact, we observe that using more sophisticated methods, like SMOTE, for upsampling also don't help with our dataset. We observe that the Positive Sampling (PS) when added standalone significantly deteriorates the ll score, F1 score and precision since it leads to an increase in the false positive rate. Adding Negative sampling(NS) standalone doesn't cause much benefit either. But adding both positive and negative sampling together leads to a boost in the precision, F1 and the ll score. We also observe that the neural network based model performs poorly compared to the decision tree based ensemble. We observe that the neural network has a very high false positive rate even after training with Negative sampling. We also include the scores computed when using a random classifier which labels any example as positive with probability 0.5 to give the readers a sense of baseline values for each of the scores.
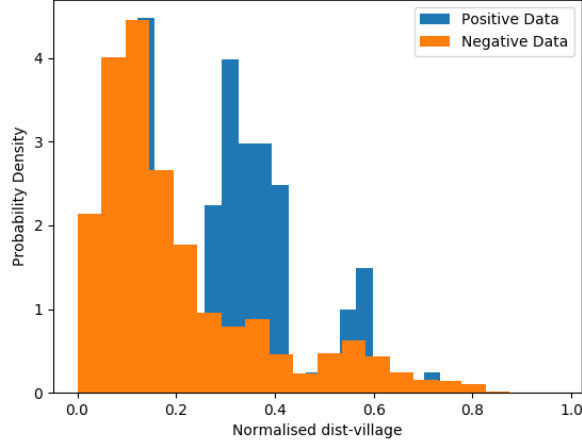
Alternate Domain Knowledge

We observed in the previous section that the model performance is very sensitive to the positive sampling strategy adopted. In fact to illustrate this we collect another type of data from the domain experts to estimate the confidence on unlabeled data. We plotted the histograms for individual features and found that in some of them there was a clear difference between positive and negative samples. The histograms for dist-village and dist-river can be seen in Figure 6 and Figure 7. Hence, we asked the experts to provide specific ranges for some of the features which they think have high correlation with the labels. We also asked for their confidence on the specific ranges provided (in line with the 4-point estimates [19] to elicit distributions). The experts felt that providing confidence for each feature was difficult since multiple other factors need to be taken into account. Thus they provided only the ranges for 11 features. A sample of 3 of those features has been shown in Table 2 for illustration purposes. We observed that the ranges for some of these features are very noisy and don't correlate well with data. For example, the range given for dist-river doesn't match with the histogram shown in the

Table 2. Feature Ranges provided by Domain experts

Snaring Threat	dist-village	dist-patrol	dist-river
low	[0,0.3]	[0,0.03]	[0.1,1]
high	[0.3,1]	[0.03,1]	[0,0.1]

Figure 6. Histogram for dist-village

Figure 7. Histogram for dist-river

Table 3. Scores when positive sampling using feature ranges

Models	LL score	Recall	Precision	F1
DT with DD	14.60	0.31	0.17	0.219
DT with DD, PSFR	6.87	0.34	0.07	0.116
DT with DD and NS	14.05	0.27	0.19	0.223
DT with DD, NS, PSFR	10.88	0.29	0.14	0.189

plot. Hence, we had to prune it down to only the features for which the ranges seemed to correlate well with data. After the pruning, we were left with ranges for 4 features namely, dist-patrol, dist-village, elevation and slope. Observations made by the experts for these 4 features seemed to correlate reasonably well with the data. But we found that even these constraints were very loose. We nevertheless tested the effectiveness of this approach for positive sampling to examine the effectiveness of this approach. We have access to the feature ranges for which the probability $P(Y = 1|X)$ is high. Let us assume that, given feature X_i and corresponding feature range, F_i, the probability, $P(Y = 1|X_i \in F_i) = p$ and $P(Y = 1|X_i \notin F_i) = q$, for the 4 features (let's assume n here for generality). We can apply bayes rule as in the Cluster Labels section, and assume similar independence assumption. Simplifying it we get,

$$P(Y = 1|X^{(j)}) = \frac{\prod_{i=1}^{n} P(Y = 1|X_i^{(j)})}{P(Y = 1)^{n-1}} \quad (2)$$

$P(Y = 1)$ can also be replaced with a constant placeholder Z. Let us assume m out of the n features fall inside the specified ranges provided. The equation above can thus be simplified,

$$P(Y = 1|X^{(j)}) = \frac{p^m q^{n-m}}{Z} \quad (3)$$

We can now calculate the confidence for any data point given the ranges for some features. In our case, we assume values of p=0.05 and q=0.01 since we couldn't get those values from the experts. Using these confidences we perform Positive sampling using feature ranges (PSFR) and report the scores in Table 3. We observe that using these features brings down the performance of the model even after feature pruning. This shows how sensitive the performance is to positive sampling. Hence positive sampling in such cases needs to be done with utmost care.

Pilot Field Test

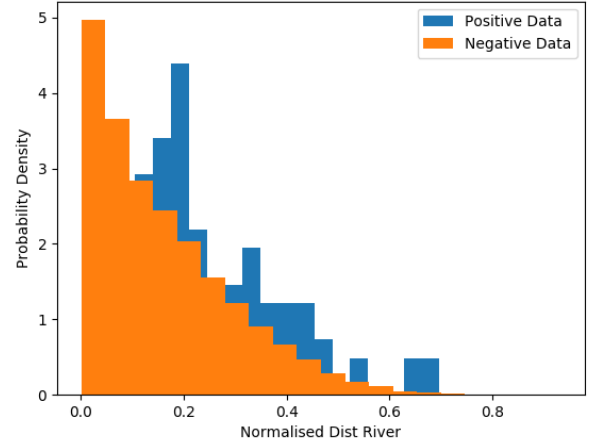In addition to evaluating our approaches using standard metrics on the dataset, we also conducted tests in the field at an early stage. In October 2017, a two-day field test was conducted in HNHR. The predictions of poaching activities were made based on a bagging ensemble of decision trees trained on 2013-2017 dataset with additional negative data points sampled from unlabeled data. The rangers selected two patrol routes that covered areas which had not been frequently patrolled but were still predicted to have poaching activities by our model. During the field test, 22 snares were found (see Figure 8), demonstrating the potential of a machine learning model enhanced with human knowledge. We expect to run larger scale tests with the best-performing model that incorporates human knowledge elicited through the clustering based questionnaire and the unlabeled data in conservation sites in Northeastern China in the next patrol season.

CONCLUSION AND FUTURE WORK

In this paper, we focus on eliciting and exploiting human knowledge to enhance the predictive analysis in wildlife poaching. The dataset in this domain has very few positive data points and suffers from uncertainty in negative labels. We designed a questionnaire to elicit knowledge from domain experts as weak labels for a set of pre-clustered data point. The weak labels are then used to estimate a prior on the threat level of each data point. The prior information is used to augment the dataset for training. We apply our approach to a basic machine learning model: a bagging ensemble of decision trees. Taking cues from the results obtained on this examplar model,

Figure 8. 22 snares found during the pilot field test in October 2017.

we expect improved performance on more complex models as well, when our approach is applied.

Through our experiments we show that making use of the human knowledge properly can lead to performance imporovement in predictive tools. We find that the performance is very sensitive to positive data augmentation. Hence incorporating the prior knowledge to augment the positive data can be very useful. But we also observed that this needs to be done carefully. Biases in the prior knowledge can affect the performance adversely in certain cases as shown in the Alternative Domain Knowledge section. We demonstrate a more principled method of collecting this prior knowledge by clustering the data points and getting weak labels over the clusters. But this can also become the bottleneck of the proposed approach. We observed that effective communication with the experts is crucial when relying on domain knowledge. But geographical/language barriers between the experts and the data scientists can come in the way, making communication very difficult.

One future direction is to explore other alternatives of collecting quantitative information from domain experts which can be effectively incorporated into machine learning models.

It is important to note that our approach is fairly generic and can be used in a variety of other settings, where the expert knowledge is costly and a large portion of the data is unlabeled. One very similar setting would be to identify and isolate feral animals in non-native regions. Applications in this domain seem very promising in light of the recent developments in visual wildlife monitoring.

REFERENCES
1. Yelena Yesha Anupam Joshi Yaacov Yesha Mary Brady Michael A. Grasso Napthali Rishe Aniket Bochare, Aryya Gangopadhyay. 2014. Integrating domain knowledge in supervised machine learning to assess the risk of breast cancer. *Int. J. Medical Engineering and Informatics* (2014).

2. Duan Biggs. 2016. Elephant poaching: Track the impact of Kenya's ivory burn. *Nature* 534, 7606 (2016), 179.

3. Leo Breiman. 1996. Bagging predictors. *Machine learning* 24, 2 (1996), 123–140.

4. N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research* (2002).

5. Shahrzad Gholami, Benjamin Ford, Fei Fang, Andrew Plumptre, Milind Tambe, Margaret Driciru, Fred Wanyama, Aggrey Rwetsiba, Mustapha Nsubaga, and Joshua Mabonga. 2017. Taking it for a test drive: a hybrid spatio-temporal model for wildlife poaching prediction evaluated through a controlled field test. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 292–304.

6. Shahrzad Gholami, Sara Mc Carthy, Bistra Dilkina, Andrew Plumptre, Milind Tambe, Margaret Driciru, Fred Wanyama, Aggrey Rwetsiba, Mustapha Nsubaga, Joshua Mabonga, and others. 2018. Adversary models account for imperfect crime data: Forecasting and planning against real-world poachers. (2018).

7. Hyeon-Woo Kang and Hang-Bong Kang. 2017. Prediction of crime occurrence from multi-modal data using deep learning. *PloS one* 12, 4 (2017), e0176244.

8. Debarun Kar, Benjamin Ford, Shahrzad Gholami, Fei Fang, Andrew Plumptre, Milind Tambe, Margaret Driciru, Fred Wanyama, Aggrey Rwetsiba, Mustapha Nsubaga, and others. 2017. Cloudy with a chance of poaching: adversary behavior modeling and forecasting with real-world poaching data. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 159–167.

9. Taewan Kim and Joydeep Ghosh. 2017. Relaxed Oracles for Semi-Supervised Clustering. *arXiv preprint arXiv:1711.07433* (2017).

10. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436.

11. Andrew M Lemieux. 2014. *Situational prevention of poaching*. Vol. 15. Routledge.

12. Jennifer F Moore, Felix Mulindahabi, Michel K Masozera, James D Nichols, James E Hines, Ezechiel Turikunkiko, and Madan K Oli. 2017. Are ranger patrols effective in reducing poaching-related threats within protected areas? *Journal of Applied Ecology* (2017).

13. Thanh H Nguyen, Arunesh Sinha, Shahrzad Gholami, Andrew Plumptre, Lucas Joppa, Milind Tambe, Margaret Driciru, Fred Wanyama, Aggrey Rwetsiba, Rob

Critchlow, and others. 2016. Capture: A new predictive anti-poaching tool for wildlife protection. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 767–775.

14. Grigoris Karakoulas Nitesh V. Chawla. 2005. Learning From Labeled And Unlabeled Data: An Empirical Study Across Techniques And Domains. *Journal of Artificial Intelligence Research* (2005).

15. Samia Saif and Douglas Craig MacMillan. 2016. Poaching, trade, and consumption of tiger parts in the Bangladesh Sundarbans. In *The Geography of Environmental Crime*. Springer, 13–32.

16. Brad Scriber. 2014. 100,000 Elephants Killed by Poachers in Just Three Years, Landmark Analysis Finds. (2014). `https://news.nationalgeographic.com/news/2014/08/140818-elephants-africa-poaching-cites-census/`.

17. Michael J Shaffer and Joseph A Bishop. 2016. Predicting and preventing elephant poaching incidents through statistical analysis, gis-based risk analysis, and aerial surveillance flight path modeling. *Tropical Conservation Science* 9, 1 (2016), 525–548.

18. Somayeh Shojaee, Aida Mustapha, Fatimah Sidi, and Marzanah A Jabar. 2013. A study on classification learning algorithms to predict crime status. *International Journal of Digital Content Technology and its Applications* 7, 9 (2013), 361.

19. Andrew SpeirsâĂŘBridge, Fiona Fidler, Marissa McBride, Louisa Flander, Geoff Cumming, and Mark Burgman. 2010. SMOTE: Synthetic Minority Over-Sampling Technique. *Risk Analysis* 16 (2010), 2098.

20. VOA. 2017. Poachers Target Africa's Lions, Vultures With Poison. (2017). `https://www.voanews.com/a/poachers-target-african-lions-vultures-with-poison/4046307.html`.

21. Ting Yu. 2007. *Incorporating Prior Domain Knowledge into Inductive Machine Learning Its implementation in contemporary capital markets*. Ph.D. Dissertation. University of Technology Sydney, Sydney, Australia.

22. Milos Hauskrecht Zhipeng Luo. 2017. Active Learning of Classification Models from Soft-Labeled Groups. LLD Workshop 2017.